8-04-04

IFW

PTO/SB/21 (04-04)

OIPE JC29

AUG 0 2 2004

PATENT & TRADEMARK OFFICE

TRANSMITTAL FORM

(to be used for all correspondence after initial filing)

		Application Number	10/626,049
		Filing Date	July 23, 2003
		First Named Inventor	Hosoya, Mutsumi
		Art Unit	2186
		Examiner Name	Unassigned
Total Number of Pages in This Submission	13	Attorney Docket Number	16869P-079400US

ENCLOSURES *(Check all that apply)*

- ☒ Fee Transmittal Form
  - ☐ Fee Attached
- ☐ Amendment/Reply
  - ☐ After Final
  - ☐ Affidavits/declaration(s)
- ☐ Extension of Time Request
- ☐ Express Abandonment Request
- ☐ Information Disclosure Statement
- ☐ Certified Copy of Priority Document(s)
- ☐ Response to Missing Parts/ Incomplete Application
  - ☐ Response to Missing Parts under 37 CFR 1.52 or 1.53
- ☐ Drawing(s)
- ☐ Licensing-related Papers
- ☒ Petition to Make Special
- ☐ Petition to Convert to a Provisional Application
- ☐ Power of Attorney, Revocation Change of Correspondence Address
- ☐ Terminal Disclaimer
- ☐ Request for Refund
- ☐ CD, Number of CD(s) _____
- ☐ After Allowance Communication to Technology Center (TC)
- ☐ Appeal Communication to Board of Appeals and Interferences
- ☐ Appeal Communication to TC **(Appeal Notice, Brief, Reply Brief)**
- ☐ Proprietary Information
- ☐ Status Letter
- ☒ Other Enclosure(s) (please identify below):
  - Return Postcard
  - 6 cited references

Remarks: The Commissioner is authorized to charge any additional fees to Deposit Account 20-1430.

SIGNATURE OF APPLICANT, ATTORNEY, OR AGENT

Firm or Individual name	Townsend and Townsend and Crew LLP Chun-Pok Leung Reg. No. 41,405
Signature	
Date	August 2, 2004

CERTIFICATE OF TRANSMISSION/MAILING

Express Mail Label: **EV 503884768 US**

I hereby certify that this correspondence is being deposited with the United States Postal Service with "Express Mail Post Office to Address" service under 37 CFR 1.10 on this date August 2, 2004 and is addressed to: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450 on the date shown below.

Typed or printed name	Joy Salvador		
Signature		Date	August 2, 2004

60275107 v1

FEE TRANSMITTAL for FY 2004

Effective 10/01/2003. Patent fees are subject to annual revision.

☐ Applicant claims small entity status. See 37 CFR 1.27

TOTAL AMOUNT OF PAYMENT ($) 130.00

Complete if Known	
Application Number	10/626,049
Filing Date	July 23, 2003
First Named Inventor	Hosoya, Mutsumi
Examiner Name	Unassigned
Art Unit	2186
Attorney Docket No.	16869P-079400US

METHOD OF PAYMENT *(check all that apply)*

☐ Check ☐ Credit Card ☐ Money Order ☐ Other ☐ None

☒ Deposit Account:

Deposit Account Number: 20-1430

Deposit Account Name: Townsend and Townsend and Crew LLP

The Director is authorized to: *(check all that apply)*

☒ Charge fee(s) indicated below ☒ Credit any overpayments

☒ Charge any additional fee(s) or any underpayment of fee(s)

☐ Charge fee(s) indicated below, **except for the filing fee** to the above-identified deposit account.

FEE CALCULATION

1. BASIC FILING FEE

Large Entity Fee Code	Large Entity Fee ($)	Small Entity Fee Code	Small Entity Fee ($)	Fee Description	Fee Paid
1001	770	2001	385	Utility filing fee	
1002	340	2002	170	Design filing fee	
1003	530	2003	265	Plant filing fee	
1004	770	2004	385	Reissue filing fee	
1005	160	2005	80	Provisional filing fee	

SUBTOTAL (1) ($)0.00

2. EXTRA CLAIM FEES FOR UTILITY AND REISSUE

			Extra Claims		Fee from below		Fee Paid
Total Claims		-** =		X		=	
Independent Claims		-** =		X		=	
Multiple Dependent				X		=	

Large Entity Fee Code	Large Entity Fee ($)	Small Entity Fee Code	Small Entity Fee ($)	Fee Description
1202	18	2202	9	Claims in excess of 20
1201	86	2201	43	Independent claims in excess of 3
1203	290	2203	145	Multiple dependent claim, if not paid
1204	86	2204	43	** Reissue independent claims over original patent
1205	18	2205	9	** Reissue claims in excess of 20 and over original patent

SUBTOTAL (2) ($)0.00

**or number previously paid, if greater; For Reissues, see above*

FEE CALCULATION (continued)

3. ADDITIONAL FEES

Large Entity Fee Code	Large Entity Fee ($)	Small Entity Fee Code	Small Entity Fee ($)	Fee Description	Fee Paid
1051	130	2051	65	Surcharge - late filing fee or oath	
1052	50	2052	25	Surcharge - late provisional filing fee or cover sheet.	
1053	130	1053	130	Non-English specification	
1812	2,520	1812	2,520	For filing a request for reexamination	
1804	920*	1804	920*	Requesting publication of SIR prior to Examiner action	
1805	1,840*	1805	1,840*	Requesting publication of SIR after Examiner action	
1251	110	2251	55	Extension for reply within first month	
1252	420	2252	210	Extension for reply within second month	
1253	950	2253	475	Extension for reply within third month	
1254	1,480	2254	740	Extension for reply within fourth month	
1255	2,010	2255	1,005	Extension for reply within fifth month	
1401	330	2401	165	Notice of Appeal	
1402	330	2402	165	Filing a brief in support of an appeal	
1403	290	2403	145	Request for oral hearing	
1451	1,510	1451	1,510	Petition to institute a public use proceeding	
1452	110	2452	55	Petition to revive – unavoidable	
1453	1,330	2453	665	Petition to revive – unintentional	
1501	1,330	2501	665	Utility issue fee (or reissue)	
1502	480	2502	240	Design issue fee	
1503	640	2503	320	Plant issue fee	
1460	130	1460	130	Petitions to the Commissioner	130
1807	50	1807	50	Petitions related to provisional applications	
1806	180	1806	180	Submission of Information Disclosure Stmt	
8021	40	8021	40	Recording each patent assignment per property (times number of properties)	
1809	770	2809	385	Filing a submission after final rejection (37 CFR § 1.129(a))	
1810	770	2810	385	For each additional invention to be examined (37 CFR § 1.129(b))	
1801	770	2801	385	Request for Continued Examination (RCE)	
1802	900	1802	900	Request for expedited examination of a design application	

Other fee (specify) _____

*Reduced by Basic Filing Fee Paid **SUBTOTAL (3)** ($)130.00

SUBMITTED BY *Complete (if applicable)*

Name (Print/Type)	Chun-Pok Leung	*Registration No. (Attorney/Agent)*	41,405	*Telephone*	650-326-2400
Signature				*Date*	August 2, 2004

WARNING: Information on this form may become public. Credit card information should not be included on this form. Provide credit card information and authorization on PTO-2038.

60275072 v1

PATENT
Attorney Docket No.: 16869P-079400US
Client Ref. No.: 310201611US1

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re application of:

MUTSUMI HOSOYA

Application No.: 10/626,049

Filed: July 23, 2003

For: HIGH-AVAILABILITY DISK CONTROL DEVICE AND FAILURE PROCESSING METHOD THEREOF AND HIGH-AVAILABILITY DISK SUBSYSTEM

Customer No.: 20350

Examiner: Unassigned

Technology Center/Art Unit: 2186

Confirmation No.: 1442

PETITION TO MAKE SPECIAL FOR NEW APPLICATION UNDER M.P.E.P. § 708.02, VIII & 37 C.F.R. § 1.102(d)

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Sir:

This is a petition to make special the above-identified application under MPEP § 708.02, VIII & 37 C.F.R. § 1.102(d). The application has not received any examination by an Examiner.

(a) The Commissioner is authorized to charge the petition fee of $130 under 37 C.F.R. § 1.17(i) and any other fees associated with this paper to Deposit Account 20-1430.

10626049

08/05/2004 HVUONG1 00000025 201430 10626049
01 FC:1460 130.00 DA

(b) All the claims are believed to be directed to a single invention. If the Office determines that all the claims presented are not obviously directed to a single invention, then Applicant will make an election without traverse as a prerequisite to the grant of special status.

(c) Pre-examination searches were made of U.S. issued patents, including a classification search and a computer database search. The searches were performed on or around June 17, 2004. The classification search covered Classes 710 (subclass 316), 711 (subclass 113), and 714 (subclasses 5 and 6), and was conducted by a professional search firm, Kramer & Amado, P.C. The computer database search was conducted on the USPTO systems EAST and WEST. The inventor further provided two references considered most closely related to the subject matter of the present application (see references #5 and #6 below), which were cited in the Information Disclosure Statement filed with the application on July 23, 2003.

(d) The following references, copies of which are attached herewith, are deemed most closely related to the subject matter encompassed by the claims:

(1) U.S. Patent Publication No. 2003/0084237 A1;

(2) Japanese Patent Publication No. 9-198308;

(3) U.S. Patent No. 5,724,542;

(4) U.S. Patent No. 5,615,330;

(5) Japanese Patent Publication No. 2002-041348; and

(6) Japanese Patent Publication No. 2002-242434.

(e) Set forth below is a detailed discussion of references which points out with particularity how the claimed subject matter is distinguishable over the references.

A. Claimed Embodiments of the Present Invention

The claimed embodiments relate to a high-availability disk control device that at no time, including at times of failure, leads to performance degradation in the storage system or to malfunctions in host applications.

Independent claim 1 recites a disk control device comprising a plurality of host interface modules configured to interface with a computer; a plurality of disk interface modules configured to interface with a storage device; a plurality of cache memory modules configured to temporarily store data read from or written to the storage device; and a switch network connecting the host interface modules, the cache memory modules, and the disk interface modules, the switch network comprising at least one switch. Each of the host interface modules is configured to execute data transfers between the computer and the cache memory modules, and each of the disk interface modules is configured to execute data transfers between the storage device and the cache memory modules. Each of the host interface modules, the disk interface modules, and the cache memory modules includes identification information providing unique identification within the switch network. The switch network includes a memory containing path information based on the identification information for data transfer paths among the host interface modules, the disk interface modules, and the cache memory modules. Each of the cache memory modules is configured to monitor failure in the cache memory module and to control changing of the path information relating to the cache memory module in the memory of the switch network.

Independent claim 6 recites a disk control device comprising a plurality of host interface modules configured to interface with a computer; a plurality of disk interface modules configured to interface with a storage device; a plurality of cache memory modules configured to temporarily store data read from or written to the storage device; a plurality of resource management modules configured to store control information relating to data transfer among the cache memory modules and the host interface modules and the disk interface modules; and a switch network connecting the host interface modules, the cache memory modules, the resource management modules, and the disk interface modules, the switch network comprising at least one switch. Each of the host interface modules is

configured to execute data transfers between the computer and the cache memory modules; and each of the disk interface modules is configured to execute data transfers between the storage device and the cache memory modules. Each of the host interface modules, the disk interface modules, the resource management modules, and the cache memory modules includes identification information providing unique identification within the switch network. The switch network includes a memory containing path information based on identification information for data transfer paths among the host interface modules, the disk interface modules, the resource management modules, and the cache memory modules. Each of the resource management modules is configured to monitor failure in the resource management module and to control changing of the path information relating to the resource management module in the memory of the switch network.

Independent claim 14 recites a disk control device comprising a plurality of host interface modules configured to interface with a computer; a plurality of disk interface modules configured to interface with a storage device; a plurality of cache memory modules configured to temporarily store data read from or written to the storage device; wherein each of the host interface modules is configured to execute data transfers between the computer and the cache memory modules, and each of the disk interface modules is configured to execute data transfers between the storage device and the cache memory modules; wherein each of the host interface modules, the disk interface modules, and the cache memory modules includes identification information providing unique identification; means for connecting the host interface modules, the cache memory modules, and the disk interface modules; and means for providing a memory containing path information based on identification information for data transfer paths among the host interface modules, the disk interface modules, and the cache memory modules, and for changing the path information for the data transfer paths in the memory, when a failure takes place in one of the cache memory modules, to avoid a failed cache memory module.

Independent claim 17 recites a disk control device comprising a plurality of host interface modules configured to interface with a computer; a plurality of disk interface modules configured to interface with a storage device; a plurality of cache memory modules configured to temporarily store data read from or written to the storage device; a plurality of resource management modules configured to store control information relating to data

transfer among the cache memory modules and the host interface modules and the disk interface modules; wherein each of the host interface modules is configured to execute data transfers between the computer and the cache memory modules, and each of the disk interface modules is configured to execute data transfers between the storage device and the cache memory modules; wherein each of the host interface modules, the disk interface modules, the resource management modules, and the cache memory modules includes identification information providing unique identification; means for connecting the host interface modules, the cache memory modules, the resource management modules, and the disk interface modules; and means for providing a memory containing path information based on identification information for data transfer paths among the host interface modules, the disk interface modules, the resource management modules, and the cache memory modules, and for changing the path information for the data transfer paths in the memory, when a failure takes place in one of the cache memory modules or the resource management modules, to avoid a failed module.

Independent claim 19 recites a failure recovery processing method for a disk control device, the method comprising providing a plurality of host interface modules configured to interface with a computer; and providing a plurality of disk interface modules configured to interface with a storage device; providing a plurality of cache memory modules configured to temporarily store data read from or written to the storage device. Each of the host interface modules is configured to execute data transfers between the computer and the cache memory modules, and each of the disk interface modules is configured to execute data transfers between the storage device and the cache memory modules. Each of the host interface modules, the disk interface modules, and the cache memory modules includes identification information providing unique identification. The method further comprises connecting the host interface modules, the cache memory modules, and the disk interface modules; providing a memory containing path information based on identification information for data transfer paths among the host interface modules, the disk interface modules, and the cache memory modules; and changing the path information for the data transfer paths in the memory, when a failure takes place in one of the cache memory modules, to avoid a failed cache memory module.

Independent claim 32 recites a disk array system for connecting to a plurality of computers via a first network. The disk array system comprises a plurality of magnetic disk devices and a disk control device connected via a second network. The disk control device comprises a plurality of host interface modules including an interface with the computers; a plurality of disk interface modules including an interface with the magnetic disk devices; and a plurality of cache memory modules connected between the plurality of host interface modules and the plurality of disk interface modules via a switch network having at least one switch. The plurality of host interface modules, the plurality of disk interface modules, and the plurality of cache memory modules each include an ID providing unique identification within the switch network. The switch includes a memory containing path information based on the IDs for data transfer paths among the host interface modules, the disk interface modules, and the cache memory modules. The disk control device comprises means for changing the path information in the memory of the switch and the IDs.

Independent claim 38 recites a disk control device comprising a plurality of host interface modules configured to interface with a computer; a plurality of disk interface modules configured to interface with a storage device; a plurality of cache memory modules configured to temporarily store data read from or written to the storage device; and a switch network connecting the host interface modules, the cache memory modules, and the disk interface modules, the switch network comprising a processor and a memory storing a program executable by the processor. Each of the host interface modules is configured to execute data transfers between the computer and the cache memory modules, and each of the disk interface modules is configured to execute data transfers between the storage device and the cache memory modules. Each of the host interface modules, the disk interface modules, and the cache memory modules includes identification information providing unique identification within the switch network. The memory of the switch network includes path information based on the identification information for data transfer paths among the host interface modules, the disk interface modules, and the cache memory modules. The program in the memory of the switch network includes a code module for changing the path information relating to the cache memory modules in response to an instruction from one of the cache memory modules upon detecting failure in the cache memory module.

One benefit that may be derived is that failure notification from the failure monitoring mechanism is analyzed by the path control mechanism and a forwarding table is controlled, thereby allowing the handling of flexible system structures. In large-scale disk control devices with multiple disk control subunits, failure information from multiple failure monitoring mechanisms can be collected by the path control mechanism to provide more reliable failure status analysis, thus providing reliable failure recovery processing.

B. Discussion of the References

None of the following references disclose or suggest providing a disk control device in which each of the host interface modules, the disk interface modules, and the cache memory modules includes identification information providing unique identification; connecting the host interface modules, the cache memory modules, and the disk interface modules; providing a memory containing path information based on identification information for data transfer paths among the host interface modules, the disk interface modules, and the cache memory modules; and changing the path information for the data transfer paths in the memory, when a failure takes place in one of the cache memory modules, to avoid a failed cache memory module.

1. U.S. Patent Publication No. 2003/0084237 A1

This reference discloses a disk array controller 1 having a plurality of disk array controlling units 1-2 and a host switch interface section 30. The disk array controlling unit 1-2 is provided with an interface (a channel interface section) 11 with the host switch interface section 30, an interface section (a disc interface section) 12 with a magnetic disc unit 5, and a cache memory section 14, in which unit a mutual connection network 21 intervenes between the channel and disc interface sections 11 and 12 and the cache memory section.14. The cache memory sections 14 of the respective disk array controlling units 1-2 are interconnected through the mutual connection network 21. It is arranged such that all of the channel interface sections 11 and the disc interface sections 12 are through the mutual connection network 21 accessible to all the cache memory sections 14. The mutual connection network 21 is arranged such that the data transfer performance thereof within a

disk array controlling unit is superior to that by way of the plurality of disk array controlling units.

2. Japanese Patent Publication No. 9-198308

This reference discloses a plurality of host computers 101, 102 connected with secondary storage devices 110, 120 through the switch 103. A cache memory 107 provided with a cache controller 104 common to the secondary storage devices 110, 120 is switch-connected to the secondary storage devices in parallel. At the time of accessing data, the host computers 101, 102 transmit data to the cache memory 107 through the switch. When a cache error occurs in the cache memory 107, a disk array management table is referred to, a port to which the secondary storage device corresponding to a logic volume in a packet is connected is specified and the secondary storage device is accessed. Data transmitted from the secondary storage device is stored in the cache memory 107 and data is transferred to the host computers 101, 102 through the switch 103.

3. U.S. Patent No. 5,724,542

This reference discloses a method of controlling disk control unit. The disk control unit 30 is equipped with first and second groups each having a channel adapter 31a for interfacing the host apparatus 11; a device adapter 32a for interfacing the direct-access storage device 40a; a resource manager 35a, equipment with a control table store, for performing control related to overall resource management and processing operations, and a service adapter 36a for executing initial microprogram loading, status monitoring processing, and malfunction recovery processing; and a cache memory 33 provided so as to be shared by the first and second groups. Data commanded from the host apparatus is written in the disk device via the cache memory or data read from the disk device is transferred to the host apparatus via the cache memory.

4. U.S. Patent No. 5,615,330

This reference discloses a method for rapidly recovering a multiprocessor data processing system from failure of a boot disk. Each data processing unit 10, 11 in the system has a private boot disk (D0-1, D0-2 with private boot disk drive 14), and at least one shared

disk (D1-1, D1-2 with shared disk drive 15). If the boot disk of one of the processing units fails, the system is temporarily reconfigured to connect a new boot disk in place of the shared disk in that processing unit. Another of the processing units is then operated to copy the contents of its own boot disk to the new boot disk.

5. Japanese Patent Publication No. 2002-041348

This reference relates to a communication pass through mechanism for providing network communication with high availability between a shared system resource and a client of the system resource. The system resource is provided with a control/processing subsystem with many peer blade processors. Ports of each blade processor are connected with each client/server network path and each client is connected with corresponding ports of each blade processor. Each blade processor is provided with a network failure detector to transfer beacon transmission with other blade processors via the corresponding blade processor port and a network path. Each blade processor redirects client communication to a failed port of other blade processor to the corresponding port of the blade processor by accepting that no beacon transmission is received from a failed port of other blade processor. As with other conventional approaches, this technique involves updating routing tables for each of the multiple processors, which renders failure handling time-consuming, prevents continuation of read/write tasks from the host computer, and can lead to performance degradation in the storage system and malfunction in application programs.

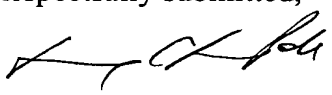
6. Japanese Patent Publication No. 2000-242434

This reference discloses a storage device system constructed to correspond to the scale or request of a computer system with the goal of easily realizing the extension of a storage device system and improvement in reliability in the future. The system 1 has a plurality of subset 10 having a storage device for holding data and a controller for controlling the storage device and switch device 20 arranged between the subsets 10 and a host 30. Each switch device 20 has a managing table for holding management information for managing the configuration of the storage device system 1. According to the management information, address information contained in the frame information outputted by the host 30 is translated and the frame information is distributed to the subsets 10. In this conventional technology, a

failure in one of the multiple disk array subsets is handled by updating routes and the like by interpreting packets within the switch and modifying requests to the failed sections so that their destinations are changed to redundant sections having equivalent functions.

(f) In view of this petition, the Examiner is respectfully requested to issue a first Office Action at an early date.

Respectfully submitted,



Chun-Pok Leung
Reg. No. 41,405

TOWNSEND and TOWNSEND and CREW LLP
Two Embarcadero Center, 8th Floor
San Francisco, California 94111-3834
Tel: 650-326-2400
Fax: 415-576-0300
Attachments
RL:rl
60263047 v1

(19) JAPANESE PATENT OFFICE

PATENT ABSTRACTS OF JAPAN

(11) Publication number: **09198308 A**

(43) Date of publication of application: **31.07.97**

(51) Int. Cl **G06F 12/08**
G06F 12/08
G06F 3/06
G06F 3/06

(21) Application number: **08023202**

(22) Date of filing: **17.01.96**

(71) Applicant: **HITACHI LTD**

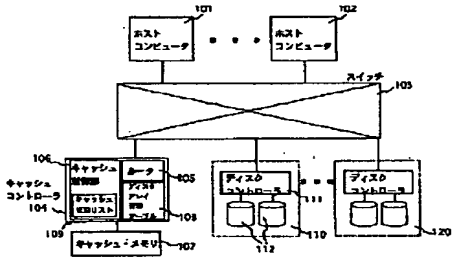
(72) Inventor: **FUJIBAYASHI AKIRA**
TAKAMOTO YOSHIFUMI

(54) **DATA STORAGE SYSTEM**

(57) Abstract:

PROBLEM TO BE SOLVED: To effectively use a cache memory and to improve data access performance in a data storage system using a switch.

SOLUTION: Plural host computers 101 and 102 are connected with secondary storage devices 110 and 120 through the switch 103. A cache memory 107 provided with a cache controller 104 common to the secondary storage devices 110 and 120 is switch-connected to the secondary storage devices in parallel. At the time of accessing data, the host computers 101 and 102 transmit data to the cache memory 107 through the switch. When a cache error occurs in the cache memory 107, a disk array management table 108 is referred to, a port to which the secondary storage device corresponding to a logic volume in a packet is connected is specified and the secondary storage device is accessed. Data transmitted from the secondary storage device is stored in the cache memory 107 and data is transferred to the host computers 101 and 102 through the switch 103.



(19)日本国特許庁（ＪＰ）　(12)公開特許公報（A）　(11)特許出願公開番号

特開平9－198308

(43)公開日　平成９年(1997)７月31日

(51)Int.Cl.⁶	識別記号	庁内整理番号	ＦＩ		技術表示箇所
Ｇ０６Ｆ　12/08		7623－5B	Ｇ０６Ｆ　12/08	Ｈ	
		7623－5B		Ｐ	
	３２０	7623－5B		３２０	
3/06	３０１		3/06	３０１Ｍ	
	３０２			３０２Ａ	

審査請求　未請求　請求項の数５　ＦＤ　（全 10 頁）

(21)出願番号　特願平8－23202

(22)出願日　平成８年(1996)１月17日

(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地

(72)発明者　藤林　昭
東京都国分寺市東恋ケ窪１丁目280番地
株式会社日立製作所中央研究所内

(72)発明者　高本　良史
東京都国分寺市東恋ケ窪１丁目280番地
株式会社日立製作所中央研究所内

(74)代理人　弁理士　笹岡　茂　（外１名）

(54)【発明の名称】　データ記憶システム

(57)【要約】

【課題】　スイッチを用いたデータ記憶システムにおいて、キャッシュメモリを有効に活用してデータアクセス性能を向上することにある。

【解決手段】　複数のホストコンピュータ101,102と２次記憶装置110,120はスイッチ103を介して接続され、２次記憶装置に共通のキャッシュコントローラ104を備えるキャッシュメモリ107が２次記憶装置と並列にスイッチに接続される。ホストコンピュータはデータをアクセスするときパケットをスイッチを介してキャッシュメモリに送出する。キャッシュメモリでキャッシュミスが生じたときは、ディスクアレイ管理テーブル107を参照してパケット内の論理ボリュームに対応する２次記憶装置が接続されているポートを特定して２次記憶装置をアクセスし、２次記憶装置から送られてくるデータをキャッシュメモリに格納し、データをスイッチを介してホストコンピュータに転送する。

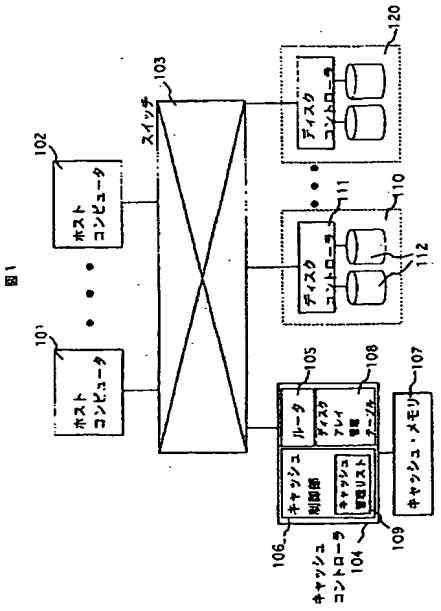


図1

1

【特許請求の範囲】

【請求項１】　複数のホストコンピュータと、複数の２次記憶装置とから構成され、該複数のホストコンピュータと該複数の２次記憶装置との間をスイッチにより接続するデータ記憶システムにおいて、該スイッチに該複数の２次記憶装置と並列に独立したキャッシュメモリを接続し、該キャッシュメモリは該複数の２次記憶装置の該スイッチを介したキャッシュメモリであることを特徴とするデータ記憶システム。

【請求項２】　請求項１記載のデータ記憶システムにおいて、

前記キャッシュメモリは、前記ホストコンピュータのデータ転送要求に対して、要求されたデータを検索し、検索の結果該データがキャッシュメモリ内にない場合には、前記スイッチを介して該データが格納されている２次記憶装置から該データを取り込み、該キャッシュメモリに格納した後にホストコンピュータに該データを転送することを特徴とするデータ記憶システム。

【請求項３】　請求項２記載のデータ記憶システムにおいて、

前記キャッシュメモリはキャッシュコントローラを備え、該キャッシュコントローラは、複数の各２次記憶装置が接続される前記スイッチのポート番号とホストコンピュータの認識している論理ボリュームを対応させたディスクアレイ管理テーブルを有し、前記検索の結果前記データがキャッシュメモリ内にない場合、該ディスクアレイ管理テーブルを参照して２次記憶装置をアクセスすることを特徴とするデータ記憶システム。

【請求項４】　請求項３記載のデータ記憶システムにおいて、

前記各ホストコンピュータのオペレーションシステムに前記ディスクアレイ管理テーブルを設け、該各ホストコンピュータは、前記複数の２次記憶装置のいずれかに直接アクセスするとき、該ディスクアレイ管理テーブルを参照して２次記憶装置をアクセスすることを特徴とするデータ記憶システム。

【請求項５】　請求項３記載のデータ記憶システムにおいて、

前記スイッチは、前記各ホストコンピュータが前記キャッシュメモリを使用するか直接前記２次記憶装置を使用するかを示すテーブルと直接前記２次記憶装置を使用する場合に用いられる前記ディスクアレイ管理テーブルと同様のテーブルからなるホスト管理テーブルを有し、前記ホストコンピュータからのアクセス要求に応じて前記ホスト管理テーブルを参照して前記キャッシュメモリまたは前記２次記憶装置に前記ホストコンピュータからのアクセスデータを転送することを特徴とするデータ記憶システム。

【発明の詳細な説明】

【０００１】

2

【発明の属する技術分野】本発明は、複数のホストコンピュータと複数の２次記憶装置をスイッチを利用して接続する構成のデータ記憶システムに関する。

【０００２】

【従来の技術】一般的なデータ処理システムは、ホストコンピュータと２次記憶装置から構成されている。２次記憶装置として使用されるのは主に磁気ディスク装置である。ここで発明者のいう磁気ディスク装置は単体のディスクドライブまたはディスクアレイを意味する。本発明の利用分野として挙げたスイッチを利用したデータ記憶システムの一例が特開平７－４４３２２号において記述されている。

【０００３】スイッチを用いてホストコンピュータと磁気ディスク装置を接続する構成のデータ記憶システムでは、ホストコンピュータおよび磁気ディスク装置の台数の変更にも柔軟に対応することが可能でスケーラビリティに優れる。このスイッチを利用したホストコンピュータと磁気ディスク装置間のデータ転送はパケット交換によって実現される。これを簡単に説明すると、ホストコンピュータはコマンドパケット（例えば、ポート番号１番：セクタ番号２：データサイズ５１２Byte：Read のような構成）を送信し、そのコマンドパケットをスイッチが解析してホストコンピュータと要求されたポート間の接続が確立したのちに、ホストコンピュータとそのポートに接続している磁気ディスク装置間でデータ転送が開始すると言う手順になる。また、従来のバススイッチと異なり、ある１組のホストコンピュータとの磁気ディスク装置の間でデータ転送が行われていても、データバスは占有されず、同時に他のホストコンピュータと磁気ディスク装置の間でもデータ転送が行える。ここで、データの読出しに関して、磁気ディスク装置では、ホストコンピュータから頻繁にアクセス要求のあるデータに対して、その都度読出し動作を行うと、メカニカルな動作を伴うためデータ転送に時間がかかるという問題がある。この問題を解決するために、磁気ディスク装置は通常キャッシュメモリを備えており、頻繁にアクセス要求のあるデータは、そのキャッシュメモリに格納することでそのデータの読出しに対するメカニカルな動作を省略し性能の向上を図っている。しかし、従来例の一つとして挙げたシステム構成では、従来のバススイッチを本発明の用いているスイッチに置き換え、ホストコンピュータとのインターフェース以下の部分はアレイコントローラが磁気ディスク装置とホストコンピュータ間のデータ転送を制御している従来の一般的構成である。この様な構成の場合には、接続された磁気ディスク装置の各々にキャッシュメモリが分散されて配置されており、それぞれのキャッシュが有効に利用されていない。

【０００４】

【発明が解決しようとする課題】従来例として挙げたデータ記憶システムおいて、接続している複数の磁気ディ

スク装置のそれぞれにキャッシュメモリが備わっている場合、性能向上の為のキャッシュメモリが、接続している磁気ディスク装置それぞれに分散して配置されることになり、それぞれのキャッシュを有効に利用できない。例えば、ある一つの磁気ディスク装置に格納されているデータに頻繁にアクセス要求が送られてくるような場合には、他の磁気ディスク装置に備わっているキャッシュメモリはほとんど利用されないことになり無駄になってしまう。

【０００５】本発明の目的は、スイッチを用いたデータ記憶システムにおいて、キャッシュメモリを有効に活用してデータアクセス性能を向上することにある。

【０００６】

【課題を解決するための手段】上記目的を達成するため、本発明は、複数のホストコンピュータと、複数の２次記憶装置とから構成され、該複数のホストコンピュータと該複数の２次記憶装置との間をスイッチにより接続するデータ記憶システムにおいて、該スイッチに該複数の２次記憶装置と並列に独立したキャッシュメモリを接続し、該キャッシュメモリは該複数の２次記憶装置の該スイッチを介したキャッシュメモリであるようにしている。また、前記キャッシュメモリは、前記ホストコンピュータのデータ転送要求に対して、要求されたデータを検索し、検索の結果該データがキャッシュメモリ内にない場合には、前記スイッチを介して該データが格納されている２次記憶装置から該データを取り込み、該キャッシュメモリに格納した後にホストコンピュータに該データを転送するようにしている。前記キャッシュメモリはキャッシュコントローラを備え、該キャッシュコントローラは、複数の各２次記憶装置が接続される前記スイッチのポート番号とホストコンピュータの認識している論理ボリュームを対応させたディスクアレイ管理テーブルを有し、前記検索の結果前記データがキャッシュメモリ内にない場合、該ディスクアレイ管理テーブルを参照して２次記憶装置をアクセスするようにしている。また、前記各ホストコンピュータのオペレーションシステムに前記ディスクアレイ管理テーブルを設け、該各ホストコンピュータは、前記複数の２次記憶装置のいずれかに直接アクセスするとき、該ディスクアレイ管理テーブルを参照して２次記憶装置をアクセスするようにしている。また、前記スイッチは、前記各ホストコンピュータが前記キャッシュメモリを使用するか直接前記２次記憶装置を使用するかを示すテーブルと直接前記２次記憶装置を使用する場合に用いられる前記ディスクアレイ管理テーブルと同様のテーブルからなるホスト管理テーブルを有し、前記ホストコンピュータからのアクセス要求に応じて前記ホスト管理テーブルを参照して前記キャッシュメモリまたは前記２次記憶装置に前記ホストコンピュータからのアクセスデータを転送するようにしている。

【０００７】

【実施例】本発明の提供するデータ記憶システムを以下に図面を示し実施例を参照して詳細に説明する。図１は、本発明によるデータ記憶システムの構成をブロック図で示したものである。１０１、１０２はホストコンピュータであり、１０３はスイッチであり、１０４はキャッシュコントローラであり、ルータ１０５とキャッシュ制御部１０６から構成される。１０７はキャッシュメモリである。１０８はキャッシュコントローラ１０４の持つディスクアレイ管理テーブルである。１０９はキャッシュ制御部１０６がキャッシュ検索時に参照するキャッシュ管理リストである。１１０、１２０は磁気ディスク装置であり、この図１中では一例としてディスクアレイ装置としており、ディスクコントローラ１１１、複数のディスク装置１１２から構成されている。磁気ディスク装置１２０も同様の構成である。この図に示すように本発明はキャッシュメモリ１０７を、複数の磁気ディスク装置１１０、１２０と共にスイッチ１０３に並列に接続することで、キャッシュメモリ１０７は前記磁気ディスク装置１１０、１２０及びホストコンピュータ１０１、１０２と、同様の１つのデバイスのごとくスイッチにより容易にアクセスできる構成となる。そして、磁気ディスク装置１１０、１２０個々にはキャッシュメモリを設けず、キャッシュメモリ１０７を磁気ディスク装置１１０、１２０が共用する。

【０００８】ここで、スイッチ１０３の内部構造の一例を図７に示す。７０１はデータのシリアル／パラレル変換を行う部分（Ｓ／Ｐ）である。７０２はスイッチ制御装置である。７０３はスイッチ機構である。ここで、スイッチ１０３のの動作の一例を簡単に説明する。図７の中に示すような構成のパケット７０４がポート１に接続されているホストコンピュータから発行される。パケットは送信先ポート番号とデータ部からなる。データ部の内容は７０５に示すように、コマンド、論理ボリューム（Ｌvol#）、ブロック番号（ＢＬＫ#）、ホスト番号（Host#）からなる。送信されてきたパケットの送信先ポートをみて、スイッチ制御装置７０２がそのポートとの接続を確立するように制御線を通じて信号を送出する。送信先ポートとの接続が確立したのち、送信先に向けてパケットまたはデータが送られる。

【０００９】次に上記システム構成においてキャッシュメモリ１０７を有効利用するために用いるデータ転送方法について述べる。図１のキャッシュ制御部１０６で行うホストコンピュータからのデータ読出し要求に対するデータ転送のフローチャートを図２に示す。先ず、ホストコンピュータはデータ読出し要求をキャッシュメモリの接続ポートに対して送る。ステップ２０１において、要求されたデータをキャッシュ制御部１０６において、送信されてきたパケット内の論理ボリュームＬvolとブロック番号ＢＬＫに基づいてキャッシュ管理リスト１０９を参照しキャッシュメモリ１０７内にデータが存在す

るかどうか検索する。次に判断ステップ２０２において、要求されたデータがキャッシュメモリ１０７内に存在する（キャッシュＨｉｔ）場合には、ステップ２０５において、要求されたデータをホストコンピュータに転送する。要求されたデータが存在しない（キャッシュＭｉｓｓ）場合には、ステップ２０３において、ルータに制御を移す。続いてステップ２０４において、ルータの制御で磁気ディスク装置から送られて来たデータをキャッシュ制御部１０６がキャッシュメモリ１０７に格納し、キャッシュ管理リストを更新する。この時、キャッシュメモリの容量一杯までデータがすでに格納されている場合は、データの追い出しが必要となる。現在追い出しの手法は様々な方法が利用されているが、ここでは最も使用頻度が少なく最も古いデータを追い出しの候補に選ぶ方法を使用することとするが、その他の方法用いたとしても本発明の実施にはなんら問題はない。そして、ステップ２０５において、ホストコンピュータに要求されたデータを転送する。

【００１０】ここで、図２のステップ２０１においてキャッシュ制御部が参照するキャッシュ管理リストの概略を図９に示す。キャッシュコントローラ内のルータが受信したパケット内の論理ボリューム番号とブロック番号がキャッシュ制御部に渡される。キャッシュメモリ内のデータはキャッシュ管理リストによりブロック単位に管理されている。現在使用中（キャッシュメモリ内にデータが存在する）論理ボリューム番号とブロック番号のリストと未使用のリストを持っており、渡された論理ボリューム番号とブロック番号が使用リスト中に存在するかどうか検索する。

【００１１】図３は前記ルータに制御を移した後の処理をフローチャートで示したものである。ここで、ルータの機能は、簡略に説明すると、ホストコンピュータから転送されたパケット内のディスクコマンドを解析し、その結果に基づいて所定の磁気ディスク装置を選択しコマンドやデータをルーティングすることである。図３中のステップ３０１において、ルータは送信されてきたパケット内の論理ボリュームをみて、ディスクアレイ管理テーブル（以下、図４において説明する。）を参照し、その論理ボリュームに対応する磁気ディスク装置が接続されているポートを特定する。ステップ３０２において、特定したポートに対しパケットを送出する。ステップ３０３において、このパケットを受けた磁気ディスク装置から転送されてくるデータを受け取る。

【００１２】これまで述べてきた実施例の通り、ホストコンピュータへのデータ転送時は常時キャッシュメモリ１０７を使用することで、例えば磁気ディスク装置１０１に格納されているデータに頻繁にアクセスがあるような場合でも、他の磁気ディスク装置に格納されているデータもキャッシュメモリ１０７に格納しておくことができ、すべての磁気ディスク装置に対してキャッシュメモリ１０７が有効利用される。

【００１３】図４は前記ディスクアレイ管理テーブルである。ホストコンピュータのオペレーティングシステム（以下、ＯＳと記述する）は、この論理ボリューム（Ｌｖｏ１）とブロック（ＢＬＫ）によりデータを指定してアクセス要求を発行する。実際のデータは複数ある磁気ディスク装置内に格納されているので、その磁気ディスク装置の接続されているポートとの対応を取るためにこのディスクアレイ管理テーブルを用いる。このディスクアレイ管理テーブルをキャッシュコントローラ１０４が持ち、前記実施例のようにホストコンピュータへのデータ転送時に常時キャッシュメモリ１０７を利用することで、ホストコンピュータのＯＳは論理ボリュームを用いてデータアクセス要求を発行することになり、複数の磁気ディスク装置が接続されていることとその内部のデータ配列の構成をユーザーに意識させず、単一の磁気ディスク装置の使用環境を提供する。

【００１４】図５では本発明の提供するシステムにおいて、実際のデータ転送の一例を簡略に示している。ここでは、ホストコンピュータ１０１からデータ１への読出し要求が発行され、ホストコンピュータ１０２からデータ２への読出し要求が発行されている場合を考える。データ１は前記キャッシュメモリ内に現在格納されており、データ２はこのキャッシュメモリ内に無く磁気ディスク装置１２０に格納されている。ホストコンピュータ１０１からのデータ１の読出し要求を図５中に示すような構成のパケット５０１として送信する。図２で示したフローチャートに従い、先ず最初に上記キャッシュメモリ１０７内が検索される。データ１はキャッシュＨｉｔするので、キャッシュメモリ１０７からホストコンピュータ１０１に転送される。同様にホストコンピュータ１０２はパケット５０２を送信して、上記キャッシュメモリ１０７内が検索されるがキャッシュＭｉｓｓとなり、キャッシュコントローラ内のルータが、図４で説明したディスクアレイ管理テーブルを参照してデータ２の格納位置を磁気ディスク装置１２０の接続されているポートと特定し、これに対してデータ２の読出し要求のパケット５０３を発行して、データを取込み、上記キャッシュメモリ１０７内に格納してから、ホストコンピュータ１０２に転送する。

【００１５】これまでに説明した実施例では接続するすべてのホストコンピュータは本発明の提供するキャッシュメモリにアクセスするシステムになっている。しかし、ホストコンピュータによっては、使用するアプリケーションの性質により、本発明の提供するキャッシュメモリより直接磁気ディスク装置にアクセスする方が性能が良いという場合がある。これに対して、以下に説明する実施例で対応する。一つの実施例は、ディスクアレイ管理テーブルをホストコンピュータのＯＳにも持たせることである。これにより、ホストコンピュータには論理

ボリュームと実際にデータの格納されている磁気ディスク装置が接続されているポートの対応が分かっているので、発行するパケットの送信先ポートにキャッシュメモリまたは磁気ディスク装置の接続されているポートを指定することで、本発明のキャッシュメモリの使用、不使用が選択可能になる（概略図を図６に示す。）

もう一つの実施例は前記スイッチにホストコンピュータのアクセス要求の管理を行う構成管理テーブルを持たせることである。図８にその概略図を示す。このホスト管理テーブル８０１は、スイッチ１０３内のスイッチ制御装置７０２に接続され、送信されてきたパケット内のホスト番号とホスト管理テーブル８０１を照合して、キャッシュメモリ１０７を使用するホストコンピュータと使用しないホストコンピュータを判断し、ホストコンピュータの発行する転送要求先をキャッシュメモリ１０７か、要求するデータの格納されている磁気ディスク装置かに振り分ける。例えば、ホストコンピュータの発行するアクセス要求をホスト管理テーブル８０１を参照してキャッシュメモリを使用しないホストである場合、ホスト管理テーブルはディスクアレイ管理テーブルと同様のテーブルを持っており、これを参照して接続先ポートを、指定された論理ボリュームに対応する磁気ディスク装置が接続しているポートに指定する。上記２通りの実施例により、本発明の提供するキャッシュメモリの使用・不使用の選択性を有するシステムを実現できる。なお、上記実施例では、ホストコンピュータ、磁気ディスク装置、キャッシュメモリの接続のためにスイッチによる接続方式を用いているが、スイッチによる接続方式に代えてループ状の接続方式を用いることも可能である。

【００１６】

【発明の効果】本発明により、スイッチを用いたデータ記憶システムにおいて、キャッシュメモリを有効に活用してデータアクセス性能が向上する。さらに、ホストコンピュータの使用するアプリケーションによってキャッシュメモリを使用するか、キャッシュメモリを不使用として直接磁気ディスク装置にアクセスするかの選択性を有する効率のよいデータ記憶システムを可能とする。

【図面の簡単な説明】

【図１】本発明によるキャッシュメモリを備えるスイッチを利用したデータ記憶システムの構成を示すブロック図である。

【図２】ホストコンピュータからの読出し要求に対するキャッシュ制御部の処理のフローチャートを示す図である。

【図３】ルータがデータをキャッシュメモリに格納する場合の処理のフローチャートを示す図である。

【図４】ディスクアレイ管理テーブルの一例を示す図である。

【図５】本発明における実際のデータの流れの一例を示した概略図である。

【図６】ディスクアレイ管理テーブルをホストコンピュータのＯＳに持たせた場合の本発明のデータ記憶システムの構成を示すブロック図である。

【図７】スイッチの構造の概略図である。

【図８】スイッチがホスト管理テーブルを備える場合の本発明のデータ記憶システムの構成を示すブロック図である。

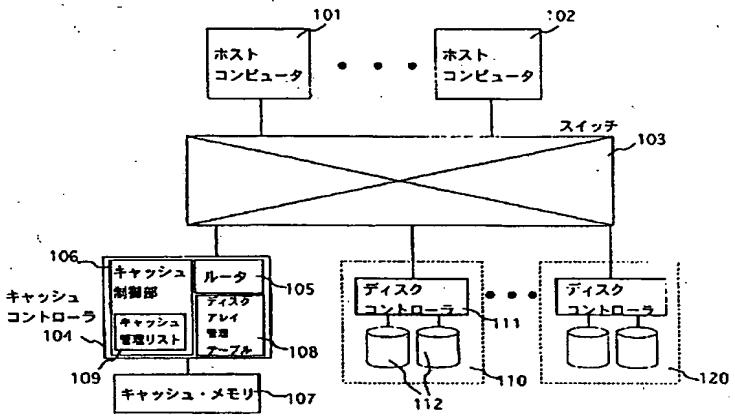
【図９】キャッシュ管理リストの概略図である。

【符号の説明】

１０１、１０２　ホストコンピュータ
１０３　スイッチ
１０４　キャッシュコントローラ
１０５　ルータ
１０６　キャッシュ制御部
１０７　キャッシュメモリ
１０８　ディスクアレイ管理テーブル
１１０、１２０　磁気ディスク装置
１１１　ディスクコントローラ
１１２　ディスク装置
７０１　Ｓ／Ｐ
７０２　スイッチ制御装置
７０３　スイッチ機構
８０１　ホスト管理テーブル

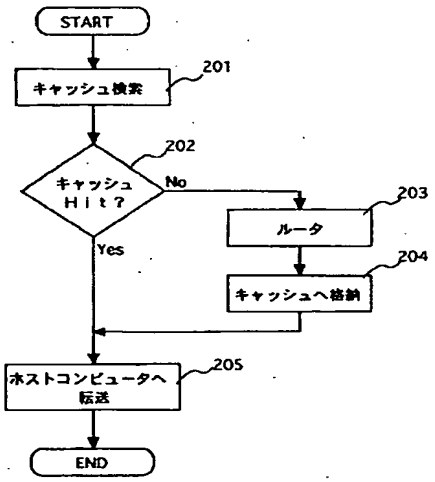
【図１】

図１



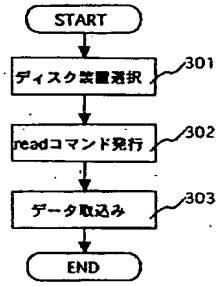
【図２】

図２



【図３】

図３



【図４】

図４

論理ボリューム番号	接続先ポート番号
0	ポート1
1	ポート1
2	ポート2
3	ポート2
4	ポート3
…	…

401　402

ディスクアレイ管理テーブル

【図５】

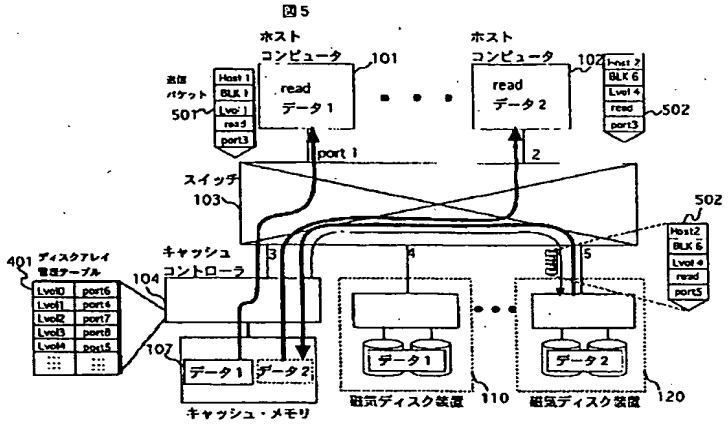


図5
ホスト
コンピュータ
101
read
データ１
ホスト
コンピュータ
102
read
データ２
502
501
port 1
2
スイッチ
103
502
401
ディスクアレイ
管理テーブル
キャッシュ
コントローラ
104
3
4
5
107
データ１
データ２
キャッシュ・メモリ
データ１
データ２
磁気ディスク装置
110
磁気ディスク装置
120

【図６】

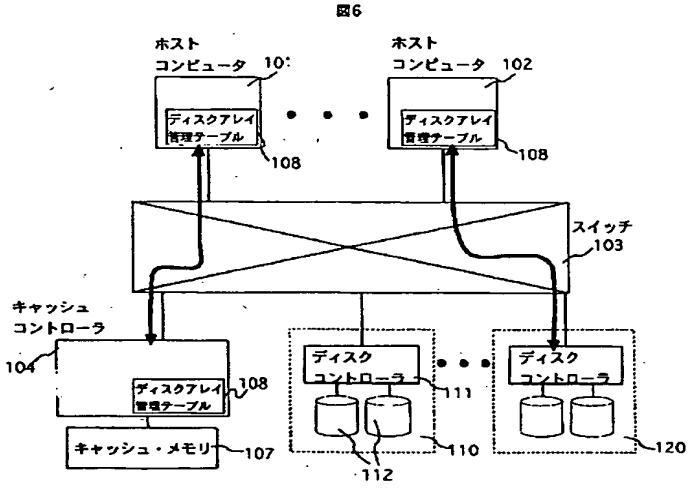
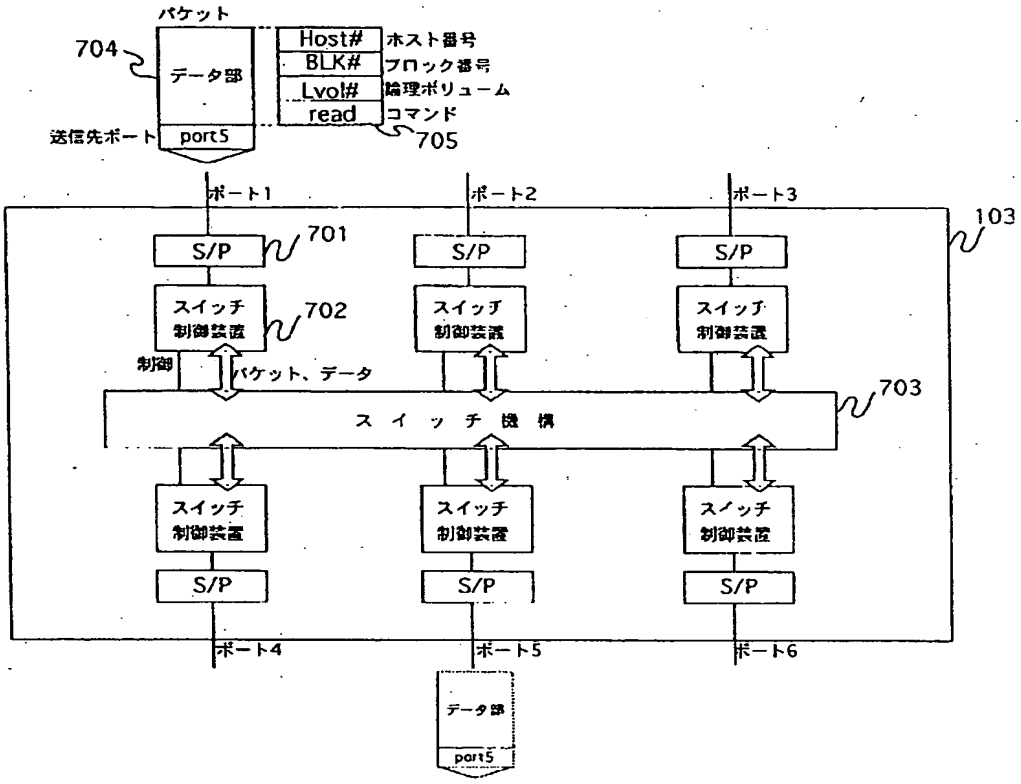


図6
ホスト
コンピュータ
101
ホスト
コンピュータ
102
ディスクアレイ
管理テーブル
108
ディスクアレイ
管理テーブル
108
スイッチ
103
キャッシュ
コントローラ
104
ディスクアレイ
管理テーブル
108
キャッシュ・メモリ
107
ディスク
コントローラ
111
ディスク
コントローラ
110
120
112

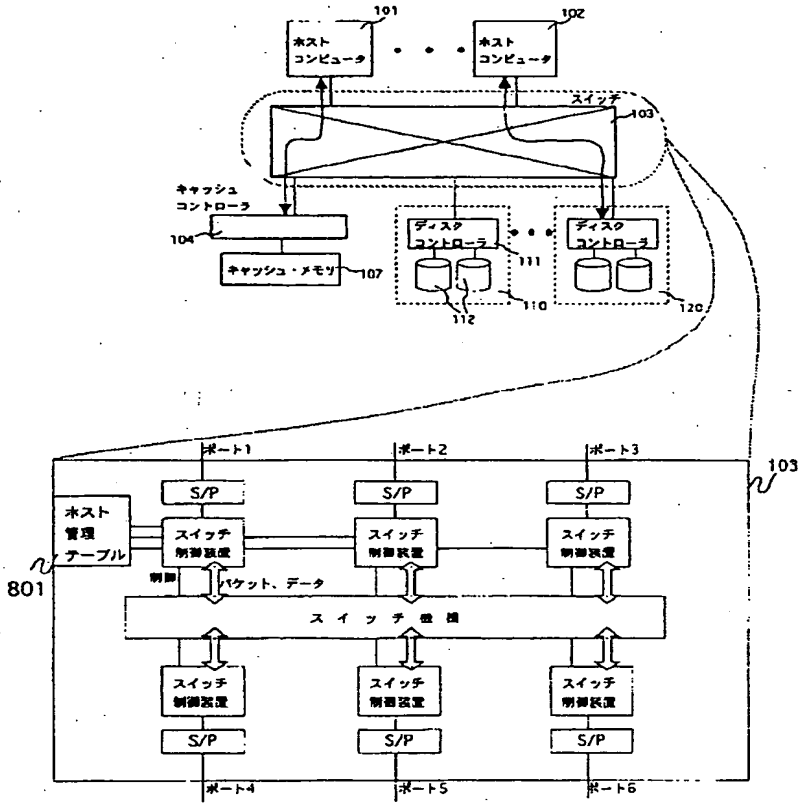
【図7】

図7


パケット
704
データ部
送信先ポート
port5
Host#
BLK#
Lvol#
read
ホスト番号
ブロック番号
論理ボリューム
コマンド
705
ポート1
ポート2
ポート3
103
S/P
701
スイッチ制御装置
702
制御
パケット、データ
703
スイッチ機構
ポート4
ポート5
ポート6
データ部
port5

【図8】

図8

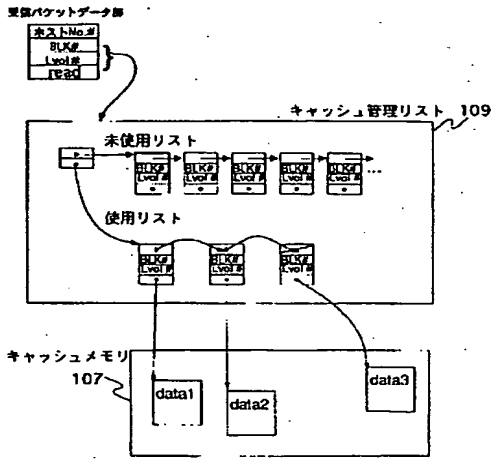


ホスト管理テーブル

host0	use	Lvol0	port4
host1	no use	Lvol1	port4
host2	use	Lvol2	port5
host3	no use	Lvol3	port5
host4	use	Lvol4	port6
host5	use	Lvol5	port6
⋮	⋮	⋮	⋮

【図９】

図9


受信パケットデータ部
ホストNo.#
BLK#
Lvol#
Read
キャッシュ管理リスト
109
未使用リスト
使用リスト
キャッシュメモリ
107
data1
data2
data3

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2002-041348
(43)Date of publication of application : 08.02.2002

(51)Int.Cl. G06F 12/00
G06F 13/00

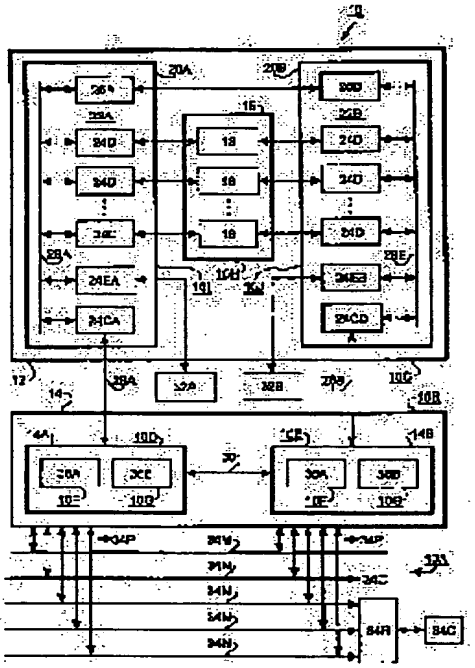
(21)Application number : 2001-155798
(22)Date of filing : 24.05.2001

(71)Applicant : EMC CORP
(72)Inventor : SCOTT JOHN A
JONES JAMES GREGORY

(30)Priority
Priority number : 2000 579428 Priority date : 26.05.2000 Priority country : US

(54) COMMUNICATION PASS THROUGH SHARED SYSTEM RESOURCE TO PROVIDE COMMUNICATION WITH HIGH AVAILABILITY, NETWORK FILE SERVER AND ITS METHOD

(57)Abstract:
PROBLEM TO BE SOLVED: To provide communication pass through mechanism to provide network communication with high availability between a shared system resource and a client of the system resource.
SOLUTION: The system resource is provided with a control/processing sub- system with many peer blade processors. Ports of each blade processor are connected with each client/server network path and each client is connected with corresponding ports of each blade processor. Each blade processor is provided with a network failure detector to transfer beacon transmission with other blade processors via the corresponding blade processor port and a network path. Each blade processor redirects client communication to a failed port of other blade processor to the corresponding port of the blade processor by accepting that no beacon transmission is received from a failed port of other blade processor.



LEGAL STATUS

[Date of request for examination] 24.05.2001

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2002－41348
（P2002－41348A）
(43)公開日　平成14年２月８日(2002.2.8)

(51)Int.Cl.'	識別記号	ＦＩ	テーマコード'(参考)
Ｇ０６Ｆ　12/00	５４５	Ｇ０６Ｆ　12/00	５４５Ａ　５Ｂ０８２
13/00	３０１	13/00	３０１Ｐ　５Ｂ０８３
	３５１		３５１Ｍ　５Ｂ０８９

審査請求　有　　請求項の数８　ＯＬ　（全 25 頁）

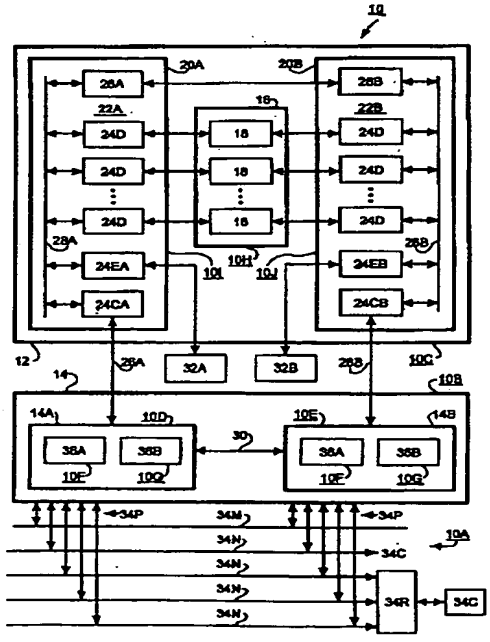
(21)出願番号　特願2001－155798(P2001－155798)
(22)出願日　平成13年５月24日(2001.5.24)
(31)優先権主張番号　０９／５７９４２８
(32)優先日　平成12年５月26日(2000.5.26)
(33)優先権主張国　米国（ＵＳ）

(71)出願人　500131642
イーエムシー　コーポレイション
アメリカ合衆国　マサチューセッツ州　01748　ホプキントン　サウス　ストリート　171
(72)発明者　ジョン　エー　スコット
アメリカ合衆国　ノースカロライナ州　27513　キャリー　トラファルガー　レーン　102
(74)代理人　100082500
弁理士　足立　勉

最終頁に続く

(54)【発明の名称】　可用性が高い通信を提供する通信パススルー共有システムリソース、ネットワークファイルサーバ及び方法

(57)【要約】
【課題】　共有システムリソースとシステムリソースのクライアントとの間で可用性の高いネットワーク通信を提供する通信パススルー機構を提供する。
【解決手段】　システムリソースは、多数のピアブレイドプロセッサを備えた制御／処理サブシステムを備える。各ブレイドプロセッサのポートは、各クライアント／サーバネットワークパスに接続され、各クライアントは、各ブレイドプロセッサの対応するポートに接続されている。各ブレイドプロセッサは、対応するブレイドプロセッサポート及びネットワークパスを介して他のブレイドプロセッサとビーコン伝送をやりとりするネットワーク故障検出器を備える。各ブレイドプロセッサは、他のブレイドプロセッサの故障したポートからビーコン伝送を受領できなかったことを受けて、他のブレイドプロセッサの故障したポートへのクライアント通信をブレイドプロセッサの対応するポートへリダイレクトする。



【特許請求の範囲】

【請求項１】　複数のクライアント／サーバ通信パスを含むネットワークを介してシステムリソースと通信するクライアントにシステムリソースサービスを提供するシステムリソースが、

システムリソース操作を実行するためのシステムリソースサブシステムと、

制御／処理サブシステムとを備え、制御／処理サブシステムが

多数のピアブレイドプロセッサを備え、各ブレイドプロセッサが各クライアント／サーバネットワーク通信パスに接続されたポートを備えるとともに各クライアントが各ブレイドプロセッサの対応するポートに接続され、

各ブレイドプロセッサが、

各クライアントの通信ルートを決定する通信ルーティングテーブルを備えた、ブレイドプロセッサとクライアントとの間の通信操作をサポートするネットワーク機構と、

ブレイドプロセッサ及びシステムリソースサブシステム間の通信とブレイドプロセッサ間の相互プロセッサ通信リンクとを提供する相互プロセッサ通信プロセッサと、

通信モニタリング機構とを備え、通信モニタリング機構が、

ブレイドプロセッサの対応するポートに接続するネットワーク通信パスを介して別のブレイドプロセッサとビーコン伝送をやりとりするためのネットワーク故障検出器と、

他のブレイドプロセッサの故障したポートからのビーコン伝送を受領できなかった際、ネットワーク故障検出器に応えて、そのブレイドプロセッサの対応するポートに、故障したポートへのクライアント通信をリダイレクトするクライアントへのリダイレクションメッセージを送信するための応答ジェネレータと、

応答ジェネレータの操作に応えて、リダイレクションメッセージに対応するように通信ルーティングテーブルを修正し、相互プロセッサ通信リンクを介して他のブレイドプロセッサとのクライアント通信をルーティングするためのパスマネージャとを備えるシステムリソース。

【請求項２】　各ブレイドプロセッサが、さらに、

別のブレイドプロセッサとの相互プロセッサ通信リンクの故障を検出し、

通信ルーティングテーブルを読み取ってそのブレイドプロセッサと他のブレイドプロセッサとの対応するポート間の機能するネットワーク通信パスを選択し、

通信ルーティングテーブルを修正して選択された機能するネットワーク通信パスを介した相互プロセッサ通信リンクを介して相互プロセッサ通信をリダイレクトするための、

相互ブレイド通信モニタを備えることを特徴とする請求項１に記載のシステムリソース。

【請求項３】　複数のクライアント／サーバ通信パスを含むネットワークを介してシステムリソース及びシステムリソースと通信するクライアント間で高可用性を備えた通信を提供する、故障に耐性がある共有システムに使用される通信パススルー機構であって、通信パススルー機構が、

システムリソース操作を実行するためのシステムリソースサブシステムと、

多数のピアブレイドプロセッサを備えた制御／処理サブシステムとを備え、各ブレイドプロセッサが各クライアント／サーバネットワーク通信パスに接続されたポートを備えるとともに各クライアントが各ブレイドプロセッサの対応するポートに接続され、

各ブレイドプロセッサが、

各クライアントの通信ルートを決定する通信ルーティングテーブルを備えた、ブレイドプロセッサとクライアントとの間の通信操作をサポートするネットワーク機構と、

ブレイドプロセッサ及びシステムリソースサブシステム間の通信とブレイドプロセッサ間の相互プロセッサ通信リンクとを提供する相互プロセッサ通信プロセッサと、

通信モニタリング機構とを備え、通信モニタリング機構が、

ブレイドプロセッサの対応するポートに接続するネットワーク通信パスを介してブレイドプロセッサと別のブレイドプロセッサとの間でビーコン伝送をやりとりするためのネットワーク故障検出器と、

他のブレイドプロセッサの故障したポートからのビーコン伝送を受領できなかった際、ネットワーク故障検出器に応えて、そのブレイドプロセッサの対応するポートに、故障したポートへのクライアント通信をリダイレクトするクライアントへのリダイレクションメッセージを送信するための応答ジェネレータと、

応答ジェネレータの操作に応えて、リダイレクションメッセージに対応するように通信ルーティングテーブルを修正し、相互プロセッサ通信リンクを介して他のブレイドプロセッサとのクライアント通信をルーティングするためのパスマネージャとを備えることを特徴とする通信パススルー機構。

【請求項４】　各ブレイドプロセッサが、さらにブレイドプロセッサと別のブレイドプロセッサとの間の相互プロセッサ通信リンクの故障を検出し、

通信ルーティングテーブルを読み取ってそのブレイドプロセッサと他のブレイドプロセッサとの対応するポート間の機能するネットワーク通信パスを選択し、

通信ルーティングテーブルを修正して選択された機能するネットワーク通信パスを介して相互プロセッサ通信をリダイレクトするための、

相互ブレイド通信モニタを備えることを特徴とする請求項３に記載のシステムリソース。

【請求項５】複数のクライアント／サーバ通信パスを含むネットワークを介してファイルサーバ及びファイルサーバのクライアント間で高可用性を備えた通信を提供する通信パススルー機構を備えた、故障に耐性があるネットワークサーバであって、ネットワークサーバが、
クライアントファイルシステム共有資源を保存するための記憶サブシステムと、
多数のピアブレイドプロセッサを備えた制御／処理サブシステムとを備え、各ブレイドプロセッサが各クライアント／サーバネットワーク通信パスに接続されたポートを備えるとともに各クライアントが各ブレイドプロセッサの対応するポートに接続され、
各ブレイドプロセッサが、
各クライアントの通信ルートを決定する通信ルーティングテーブルを備えた、ブレイドプロセッサとクライアントとの間の通信操作をサポートするネットワーク機構と、
ブレイドプロセッサ及び記憶サブシステム間の通信とブレイドプロセッサ間の相互プロセッサ通信リンクとを提供する相互プロセッサ通信プロセッサと、
通信モニタリング機構とを備え、通信モニタリング機構が、
ブレイドプロセッサの対応するポートに接続するネットワーク通信パスを介してブレイドプロセッサと別のブレイドプロセッサとの間でビーコン伝送をやりとりするためのネットワーク故障検出器と、
他のブレイドプロセッサの故障したポートからのビーコン伝送を受領できなかった際、ネットワーク故障検出器に応えて、そのブレイドプロセッサの対応するポートに、故障したポートへのクライアント通信をリダイレクトするクライアントへのリダイレクションメッセージを送信するための応答ジェネレータと、
応答ジェネレータの操作に応えて、リダイレクションメッセージに対応するように通信ルーティングテーブルを修正し、相互プロセッサ通信リンクを介して他のブレイドプロセッサとのクライアント通信をルーティングするためのパスマネージャとを備えることを特徴とするネットワークファイルサーバ。
【請求項６】　各ブレイドプロセッサが、さらに、
ブレイドプロセッサと別のブレイドプロセッサとの間の相互プロセッサ通信リンクの故障を検出し、
通信ルーティングテーブルを読み取ってそれらのブレイドプロセッサのポート間の機能するネットワーク通信パスを選択し、
通信ルーティングテーブルを修正して選択された機能するネットワーク通信パスを介して相互プロセッサ通信をリダイレクトするための、
相互ブレイド通信モニタを備えることを特徴とする請求項５のファイルサーバ。
【請求項７】複数のクライアント／サーバ通信パスを含むネットワークを介してシステムリソースと通信するクライアントにシステムリソースサービスを提供するリソースシステムにおいて、システムリソースとシステムリソースのクライアントとの間で高可用性を備えた通信を提供する方法であって、システムリソースが、システムリソース操作を実行するためのシステムリソースサブシステムと多数のピアブレイドプロセッサを備えた制御／処理サブシステムとを備え、各ブレイドプロセッサが各クライアント／サーバネットワーク通信パスに接続されたポートを備えるとともに各クライアントが各ブレイドプロセッサの対応するポートに接続され、各ブレイドプロセッサが、ブレイドプロセッサとクライアントとの間の通信操作をサポートするネットワーク機構と、ブレイドプロセッサとシステムリソースサブシステムとの間の通信を提供する相互プロセッサ通信プロセッサとを備え、方法が、
ブレイドプロセッサにおいて、
ブレイドプロセッサの対応するポートを接続するネットワーク通信パスを介して他のブレイドプロセッサとビーコン伝送をやりとりすることにより別のブレイドプロセッサの通信操作をモニタリングするステップと、
他のブレイドプロセッサの故障したポートからのビーコン伝送を受領できなかった際、そのブレイドプロセッサの対応するポートに、故障したポートへのクライアント通信をリダイレクトするクライアントへのリダイレクションメッセージを送信するステップと、
相互プロセッサ通信リンクを介して他のブレイドプロセッサとのリダイレクトされたクライアント通信をルーティングするステップとを備えた方法。
【請求項８】　システムリソースとシステムリソースのクライアントとの間で高可用性を備えた通信を提供する請求項７の方法であって、方法が、さらに、
ブレイドプロセッサにおいて、
ブレイドプロセッサと別のブレイドプロセッサとの間の相互プロセッサ通信リンクの故障を検出するステップと、
ブレイドプロセッサと他のブレイドプロセッサとの対応するポートの間の機能するネットワーク通信パスを選択するステップと、
選択された機能するネットワーク通信パスを介して相互プロセッサ通信をリダイレクトするステップとを備えることを特徴とする方法。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、ネットワークファイルサーバのような、故障に耐性がありレイテンシが低い共有システムリソースにおける高レベルトランザクションロギング機構のためのシステム及び方法、特に、多重サーバシステムリソースにおいて利用されるクロスサーバ高レベルミラードトランザクションロギング機構に

関する。

【０００２】

【従来の技術】コンピュータシステムにおいて絶えず問題となるのは、安全で故障に耐性があるリソースを提供すること、例えばコンピュータシステムとコンピュータシステムのクライアントまたはユーザとの間の通信が故障の際にも維持されるような通信リソース、そして故障の際にデータが失われずかつ損失を被ることなく回復または再構築されるようなデータ記憶リソースを提供することである。この問題は、システムデータ記憶機器のように、通常、共有リソースが１つ以上のシステムリソース、例えば、多数のクライアント間で共有され、システムネットワークを通じてアクセスされるファイルサーバから構成されるネットワークシステムにおいては特に解決するのが難しい。共有リソースにおける故障、例えば、ファイルサーバのデータ記憶機能における故障、あるいはファイルサーバのクライアントとファイルサーバによりサポートされるクライアントファイルシステムとの間の通信における故障は、システム全体の故障に発展する恐れがある。この問題は、データ量及び通信量と、ファイルサーバのような共有リソースによってサポートされるデータトランザクション数とが単一クライアントシステム内におけるそれらに比べて著しく大きいという点で特に厳しいものであり、その結果、リソース、データトランザクション、クライアント／サーバ通信における複雑さを著しく大きくしてしまう。この複雑さの増大は故障の可能性を増大させ、故障からの回復をより難しくする。さらに、その問題は、故障が、ディスクドライブや制御プロセッサ、あるいはネットワーク通信のような、数多くのリソースコンポーネントまたは関連する機能のどれにでも起こりうるという点で多次元的である。また、共有リソース通信及びサービスが１つ以上のコンポーネントに故障が起きても利用可能であり続け、さらに、リソースの操作が、完了した操作及びトランザクションと、故障が起きたときに実行されていた操作及びトランザクションとの両方について保存され回復されることが望ましい。

【０００３】ネットワークファイルサーバシステムを従来技術の共有システムリソースの典型的な例として考えると、従来技術のファイルサーバシステムは、クライアント／サーバ通信及びファイルサーバのファイルトランザクション機能においてフォールトトレランスを達成するため、そしてデータの回復または再構築のために数多くの方法を採用してきた。これらの方法は、リダンダンシ、すなわち、複写システムエレメントの供給と、故障したエレメントの複写エレメントへの置き換え、あるいは失われた情報を再構築するのに用いられる情報の複写コピーの作成とに基づくものが代表的である。

【０００４】例えば、従来技術の多くのシステムが、データ及びファイルトランザクションの保存及び回復に業界標準のＲＡＩＤ技術を組み込んでいる。ＲＡＩＤ技術は、予備のデータ及びエラー訂正情報を複数のディスクドライブの予備アレイに渡って分散する一群の方法である。故障したディスクドライブは予備のドライブに置換され、故障したディスクのデータは予備のデータ及びエラー訂正情報から再構築される。従来技術のその他のシステムは、クライアント／ファイルサーバ通信及びクライアント／クライアントファイルシステム通信の信頼性及び可用性を高めるために、故障した通信パスまたはファイルプロセッサからの通信またはファイル処理を同等の並列パスまたはプロセッサに切り換える適当なスイッチング機能を備えた多重複写式並列通信パスまたは多重複写式並列処理ユニットを採用している。しかしながら、これらの方法は、主要な通信パス及び処理パスの複写、そして、故障したエレメントを機能するエレメントに交換するのに複雑な管理及び同期機構を必要とするので、システムリソースに多額の費用がかかる。また、これらの方法により、故障の際にサービス及び機能が継続して実行され、例えばＲＡＩＤの利用により、完了したデータトランザクション、すなわち、ディスク上の固定記憶装置にコミットされたトランザクションが回復または再構築されるが、これらの方法は、トランザクションの実行中の故障により失われたトランザクションの再構築または回復をサポートしない。

【０００５】この結果、従来技術の別の方法においては、トランザクションの実行中に起きる故障により失われたトランザクションの回復及び再構築のために情報リダンダンシが利用される。これらの方法には、キャッシング、トランザクションロギング、ミラーリングが含まれる。キャッシングとは、固定記憶装置、すなわちディスクドライブへのデータの移動により固定記憶装置にデータトランザクションがコミットされるまで、あるいはデータトランザクションが固定記憶装置から読み取られて受け手に送られるまで、固定記憶装置への及びそれからのデータフローパスのメモリ中にデータを一時的に記憶することである。トランザクションロギング、あるいはジャーナリングとは、データトランザクションが固定記憶装置にコミットされるまで、すなわちファイルサーバにおいて完了されるまで、一時的にデータトランザクションを記述する情報、すなわち要求されたファイルサーバ操作を記憶し、さらに、記憶された情報から失われたデータトランザクションを再構築または再実行することである。ミラーリングは、多くの場合キャッシングまたはトランザクションロギングと共に用いられ、基本的に、キャッシュまたはトランザクションログの記録がファイルプロセッサで生成されるときに、例えば、別のプロセッサのメモリまたは固定記憶空間にキャッシュまたはトランザクションログの内容のコピーを保存することである。

【０００６】しかしながら、キャッシング、トランザク

ションロギング、ミラーリングは、あまり満足のいくものではない。なぜなら、それらは多くの場合システムリソースを高額にし、キャッシング、トランザクションロギング、ミラーリング機能及びそれに続くトランザクションの回復操作を行うために複雑な管理及び同期操作と、機構とを必要とし、著しくファイルサーバのレイテンシ、すなわちファイルトランザクションを完了するのに要する時間を増加するからである。また、キャッシング及びトランザクションロギングは、キャッシング及びロギング機構が存在するプロセッサの故障に弱いこと、また、ミラーリングがキャッシュまたはトランザクションログの内容の損失問題への解決である一方で、ミラーリングは、キャッシングまたはトランザクションロギングと同様の欠点を有することに注意しなければならない。これらの問題は、キャッシングと、特にトランザクションロギング及びミラーリングとがトランザクションロギングの間に莫大な量の情報の保存を必要とする点、及び、ログファイルトランザクションの再構築または再実行が、ファイルトランザクションの再構築のために、トランザクションログを分析し、再生し、ロールバックする複雑なアルゴリズムの実装を必要とする点でより複雑となる。また、これらの方法が、各データトランザクションが非常に多くの詳細で複雑なファイルシステム操作として実行されているようなより低いレベルのファイルサーバ機能で実装される場合が多いという点で、これらの問題はさらに複雑になる。その結果、抽出され保存されるべき情報量と、データあるいはデータトランザクションを抽出して保存し、データまたはデータトランザクションを回復及び再構築するために必要となる操作の数及び複雑さとは著しく増大する。

【０００７】また、これらの方法はシステムリソースを割高にし、それらの方法を管理するための複雑な管理及び同期機構を必要とする。そして、システムリソースが割高であるために、これらの方法が提供できるリダンダンシの度合いは制限されるので、システムは、多くの場合、複数のソースに起こる故障に対応できない。例えば、システムがある機能のために複写式並列プロセッサユニットまたは通信パスを設けても、両方のプロセッサユニットまたは通信パスで故障が起きればシステム全体が失われてしまう。さらに、通信及びデータの保存及び回復を保証するこれらの従来技術は、通常、互いから隔絶された状態で、そして異なるレベルまたはサブシステムで動作する。このため、通常、これらの方法は協力してまたは連動して動作するわけではなく、互いに相反して動作するかもしれず、複数の故障または連動した故障、またはいくつかの方法を組み合わせて対処する必要のある故障に対応できない。従来技術のいくつかのシステムは、この問題を解決しようと努力しているが、それには、中央統一的な調整機構、またはサブシステムと、協調操作を行い、故障を扱う機構間の衝突を避けるための互いに関連する複雑な管理及び同期機構を必要とし、そのためにまたシステムリソースにお金がかかるとともに、それ自体が故障の原因となる。

【０００８】

【発明が解決しようとする課題】本発明の目的は、これらの、そしてその他の従来技術に関連する問題への解決を提供することである。本発明は、複数のクライアント／サーバ通信パスを含むネットワークを介して、システムリソースとシステムリソースのクライアントとの間に可用性の高い通信を提供するための、故障に耐性のある共有システムリソース、例えばネットワークファイルサーバ、に使用される通信パススルー機構及び通信パススルー機構の操作方法に関する。

【０００９】

【課題を解決するための手段及び発明の効果】本発明によると、システムリソースには、システムリソース操作を実行するためのシステムリソースサブシステムと、多数のピアブレイドプロセッサを備えた制御／処理サブシステムとが含まれる。各ブレイドプロセッサは、各クライアント／サーバネットワーク通信パスに接続されたポートを備え、各クライアントは、各ブレイドプロセッサの対応するポートに接続されている。各ブレイドプロセッサは、各クライアントの通信ルートを決定する通信ルーティングテーブルを備えた、ブレイドプロセッサとクライアントとの間の通信操作をサポートするネットワーク機構と、ブレイドプロセッサとシステムリソースサブシステムとの間で通信を提供する相互プロセッサ通信プロセッサと、ブレイドプロセッサ間の相互プロセッサ通信リンクとを備える。各ブレイドプロセッサは、さらに、ブレイドプロセッサの対応するポートを接続するネットワーク通信パスを介して、別のブレイドプロセッサとビーコン伝送をやりとりするためのネットワーク故障検出器を備えた通信モニタリング機構を備える。各ブレイドプロセッサは、他のブレイドプロセッサの故障したポートからビーコン伝送を受領できなかった際、ネットワーク故障検出器に応えて、ブレイドプロセッサの対応するポートへ反対側のブレイドプロセッサの故障したポートへのクライアント通信をリダイレクトするクライアントへのリダイレクションメッセージを送信するための応答ジェネレータを備える。ブレイドプロセッサのパスマネージャは、応答ジェネレータの操作に応えて、リダイレクションメッセージに対応するように通信ルーティングテーブルを修正し、相互プロセッサ通信リンクを介して他のブレイドプロセッサへ他のブレイドプロセッサの故障したポートへのクライアント通信をルーティングする。

【００１０】本発明のさらなる実施例において、各ブレイドプロセッサは、相互ブレイド通信モニタを備え、別のブレイドプロセッサとの相互プロセッサ通信リンクの故障を検出し、通信ルーティングテーブルを読み取って

ブレイドポート間の機能するネットワーク通信パスを選択し、通信ルーティングテーブルを修正して相互プロセッサ通信リンクからの相互プロセッサ通信を選択された機能するネットワーク通信パスへリダイレクトする。

【００１１】

【発明の実施の形態】本発明の前述及びその他の目的、特徴、利点を、添付の図を参照しながら、実施例を用いて以下に説明する。

Ａ．高可用性を備えた共有リソースの概略説明（図１）

１．序論

以下に記述するように、本発明は、ネットワークシステムにおいて多数のユーザ間で共有されるファイルサーバ、通信サーバ、あるいはプリンタサーバのように、可用性の高いリソースに関するものである。本発明のリソースは、統合された協働クラスタからなる階層及びピアドメインから構成される。各ドメインは、リソースによってサポートされた機能またはサービスに不可欠な１つ以上の関連した機能を実行あるいは提供する。１つのドメインは、複数のサブドメインから構成されてもよいし、あるいは複数のサブドメインを具備していてもよい。例えば、１つ以上のドメインが、リソースとネットワーククライアントとの間で通信サービスを提供し、その他のドメインが、高レベルファイルシステム、通信、または印刷機能を実行し、その一方で、別のドメインが低レベルファイルシステム、通信及びプリント機能を実行してもよい。階層的に関連したドメインの場合、１つのドメインが別のドメインを制御するか、または、関連したより高いあるいは低いレベルの機能を実行することにより、より高いあるいは低いレベルのドメインをサポートすることができる。例えば、より高レベルのドメインは、関連した低レベルドメインがより低レベルのファイルまたは通信機能を実行する間、高レベルのファイルまたは通信機能を実行することができる。ピアドメインは、例えばタスクの負荷を分担してある機能についてのリソース容量を増やすために、同一あるいは並列の機能を実行したり、あるいは、共に１つのドメインを構成するために中立的なサポート関係で関連するタスクまたは機能を実行することができる。さらに、他のドメインは、ある機能についてはピアドメインであったり、他の機能については階層的に関連したドメインであったりもできる。最後に、以下に説明するように、あるドメインは、他のドメインの故障処理機構とは別に独立して動作するが、高レベルのリソース可用性を達成するために協調的に動作する故障処理機構を備える。

【００１２】本発明は、例えば、そして以下に説明する目的で、高可用性を備えたネットワークファイルサーバ（ＨＡＮファイルサーバ）１０に実装される。この実装の形態を、本発明の実施例として以下に詳細に記述する。図１に示すように、本発明が実装されているＨＡＮファイルサーバ１０には、例えば、データジェネラルコーポレーション（Ｄａｔａ　Ｇｅｎｅｒａｌ　Ｃｏｒｐｏｒａｔｉｏｎ）のＣＬＡＲｉｉＯＮ™ファイルサーバを使用する。ＣＬＡＲｉｉＯＮ™ファイルサーバは、高い可用性を備えたファイルシステム共有資源、すなわち、記憶空間をネットワーククライアントに提供するとともに、ジャーナルファイルシステム、ネットワークフェイルオーバー能力、データのバックエンドレイド（ＲＡＩＤ）記憶装置を利用して、共有資源に書き込まれたデータに高い整合性を提供する。本実装においては、ＨＡＮファイルサーバ１０は、業界標準の共通インターネットファイルシステムプロトコル（ＣＩＦＳ）とネットワークファイルシステム（ＮＦＳ）共有資源との両方をサポートしており、ＣＩＦＳ及びＮＦＳによって使用されるようなファイルアクセス制御のための対照モデルが外からはわからないように実装されている。ＨＡＮファイルサーバ１０はまた、マイクロソフトウィンドウズＮＴ環境におけるドメインコントローラあるいはＵＮＩＸ（登録商標）環境のためのネットワークファイルシステム（ＮＦＳ）などの既存の業界標準管理データベースを統合している。

【００１３】本実装は、ゼロコピーＩＰプロトコルスタックを利用して高いパフォーマンスを提供する。そのために、ファイルシステムキャッシング方式をバックエンドＲＡＩＤ機構と緊密に統合するとともに、保存用のディスクへの書き込みを廃するために、ピア記憶プロセッサ上でミラーリングすることにより重要なデータの可用性を提供できるデュアル記憶プロセッサを使用する。以下に詳細に説明するように、本実装のＨＡＮファイルサーバ１０は、デュアルプロセッサファンクショナルマルチプロセッシングモードで動作している。このモードでは、１つのプロセッサが、クライアントとディスクに存在するファイルシステムとの間でデータを転送するための全てのネットワーク及びファイルシステム操作を実行するフロントエンドプロセッサとして働き、ネットワークスタック、ＣＩＦＳ／ＮＦＳの実装、ジャーナルファイルシステムをサポートする。第二プロセッサは、ブロック記憶プロセッサとして働き、可用性の高いＲＡＩＤ構成において管理されたひとまとまりのディスクへの及びそれからのデータの読み取り及び書き込みの全ての機能を実行する。

【００１４】本実装において、ファイルシステムは、カーネルベースのＣＩＦＳネットワークスタックを備えたジャーナル機能付きクイックリカバリファイルシステムとして実装され、第二モードでＮＦＳ操作をサポートするが、本発明によると、ファイルシステムのデータへのアクセスに高い可用性を提供するために修正を加えられている。ファイルシステムはさらに、ある記憶プロセッサ上のメモリに記憶されたデータ変更がその記憶プロセッサのハードウェアまたはソフトウェア故障の際に保存されるというデータ反映機能を使って、ネットワークク

ライアントがファイルシステムに加える全てのデータ変更を記憶することにより記憶プロセッサの損失に対する保護を提供する。ファイルシステムに対するコア内部のデータ変更の反映は、相互記憶プロセッサ通信システムを通じて達成され、これにより、一方の記憶プロセッサ上でクライアントによってＮＦＳまたはＣＩＦＳを使用して伝達されたファイルシステムへのデータ変更は、データを記憶しているネットワーククライアントに通知が返される前に、他方の記憶プロセッサにより反映され、受領確認される。このことは、最初の記憶プロセッサ上での故障の際に代わりの記憶プロセッサにデータ変更のコピーが取り込まれ、万が一故障が起きた際には、ファイルシステムが代わりの記憶プロセッサに引き継がれた後に、その変更がファイルシステムに適用されることを保証する。後述するように、この反映機構が、ファイルを追跡するために用いられるシステムメタデータを回復及び修復する基本的なファイルシステム回復機構の頂点に構築される一方で、反映機構はユーザデータを回復あるいは修復する機構を提供する。ブロック記憶サブシステムは、ＲＡＩＤ技術を使用してディスクユニットの損失に対しディスクレベルでの保護を提供する。ディスクドライブが失われると、ＲＡＩＤ機構は、代わりのドライブにデータを再構築する機構を提供し、失われたドライブなしで動作する際、そのデータへのアクセスを提供する。

【００１５】後述するように、本実装のＨＡＮファイルサーバ１０は、サーバのクライアントと、予備のコンポーネント及びデータパスを利用してサーバ上でサポートされたクライアントファイルシステムとの間で可用性の高い通信を提供し、クライアントとクライアントファイルシステムとの間の通信を維持するための通信故障処理機構を提供する。本発明のＨＡＮファイルサーバ１０はまた、ファイルトランザクション及びデータのバックアップ及び回復システムを備え、ファイルトランザクション及びデータの損失を防ぐとともに、ファイルトランザクション及びデータの回復または再構築を許容する。システムハードウェアまたはソフトウェア故障の際には、システムの生き残ったコンポーネントが故障したコンポーネントのタスクを引き継ぐ。例えば、記憶プロセッサ上のイーサネット（登録商標）ポートが１つ失われると、そのポートからのネットワークトラフィックは代わりの記憶プロセッサの別のポートによって引き継がれる。同様に、記憶プロセッサのどの部分かにその処理機能を危うくするような故障が起きたならば、全てのネットワークトラフィック及びファイルシステムが生き残った記憶プロセッサへ移転される。さらなる例では、データ及びファイルトランザクション及びバックアップ機構は、故障したコンポーネントが回復した際、故障したコンポーネントによる、あるいは対応するコンポーネントによるデータ及びファイルトランザクションの回復及び再構築を可能にするとともに、生き残ったコンポーネントが故障したコンポーネントのファイルトランザクションを引き継ぐことを可能にする。さらに、ディスクドライブが１つ失われても、そのディスクのデータへのアクセスが失われない。なぜなら、ＲＡＩＤ機構が生き残ったディスクを用いて、失われたドライブ上にあった再構築されたデータへのアクセスを提供するからである。全てのファイルサーバに影響を及ぼす停電の際には、停電の際のファイルサーバ状態が保存され、コア内部のデータは固定記憶装置にコミットされて電源が復旧すると回復される。これにより、停電前になされた全てのデータ変更が保存される。最後に、ＨＡＮファイルサーバ１０の通信そしてデータ及びファイルトランザクションの故障回復機構は、サーバの各ドメインまたはサブシステムに設けられ、互いに別々に独立して機能するが、ファイルシステム通信へのクライアントの可用性を高レベルに保ち、データ及びファイルトランザクションの損失を防いで回復を可能にするために、協調的に動作する。それにも関わらず、ＨＡＮファイルサーバ１０の故障回復機構は、故障のソースを特定して隔離するのに通常必要な複雑な機構や手続き、さらには衝突する可能性のある故障管理操作を調整し、同期させ、管理するのに通常必要な複雑な機構及び操作を必要としない。

【００１６】２．ＨＡＮファイルサーバ１０の詳細説明（図１）

図１には、データジェネラルコーポレーションのＣＬＡＲｉｉＯＮ™ファイルサーバのような、本発明が実装されうる典型的なＨＡＮファイルサーバ１０が示されている。図に示すように、ＨＡＮファイルサーバ１０は、記憶サブシステム１２と、記憶サブシステム１２を共有するデュアルコンピュートブレイド（ブレイド）１４Ａ及び１４Ｂからなる制御／プロセッササブシステム１４とを備える。コンピュートブレイド１４Ａ及び１４Ｂは、ＨＡＮファイルサーバ１０のクライアントに、ネットワークアクセス及びファイルシステム機能を提供及びサポートするために独立して動作し、相互バックアップと、ネットワークアクセス及び互いのファイルシステム機能のサポートとを提供するために協調的に動作する。

【００１７】ａ．記憶サブシステム１２（図１）

記憶サブシステム１２は、複数のハードディスクドライブ１８からなるドライブバンク１６を備える。各ディスクドライブ１８は、記憶ループモジュール２０Ａ及び２０Ｂとして示されるデュアル記憶ループモジュール２０（２０Ａ及び２０Ｂを総称して２０ともいう。以下同じ。）を通して双方向に読み取り／書き込みアクセスされる。図に示すように、記憶ループモジュール２０Ａ及び２０Ｂにはそれぞれ、ＭＵＸＢＡＮＫ２２Ａ及び２２Ｂとして示されるマルチプレクサバンク（ＭＵＸＢＡＮＫ）２２が含まれる。ＭＵＸＢＡＮＫ２２Ａ及び２２Ｂにはそれぞれ、複数のマルチプレクサ（ＭＵＸ）２４

と、ループコントローラ２６A及び２６Bとして示されるループコントローラ２６とが含まれる。各ループコントローラモジュール２０のMUX２４とループコントローラ２６とは、MUXループバス２８A及び２８Bとして示されたMUXループバス２８を介して双方向に相互接続されている。

【００１８】図に示すように、MUXBANK２２A及び２２Bにはそれぞれ、対応するディスクドライブ１８に対応して接続されているディスクドライブMUX２４（MUX２４D）が含まれる。そのため、ドライブバンク１６の各ディスクドライブ１８は、MUXBANK２２A及び２２Bのそれぞれにおいて、対応するDMUX２４Dに接続され、双方向に読み取り／書き込みされる。MUXBANK２２A及び２２Bはさらに、それぞれ、対応するコンピュートブレイド１４A及び１４Bの一方と、MUX２４CA及びMUX２４CBそれぞれを介して双方向に接続されており、コンピュートブレイド１４A及び１４Bはブレイドバス３０を介して双方向に接続されている。さらに、MUXBANK２２A及び２２Bは、それぞれ、MUX２４EA及び２４EBで示される外部ディスクアレイMUX２４を備えていてもよい。外部ディスクアレイMUX２４は、対応するMUXループバス２８A及び２８Bから双方向に接続され、外部ディスクアレイMUX（EDISKA）３２に双方向に接続されている。外部ディスクアレイMUX３２は、図において、それぞれEDISKA３２A及び３２Bとして示され、予備のあるいは代わりのディスク記憶空間を提供する。

【００１９】従って、各ディスクドライブ１８は、MUXBANK２２AのMUX２４及びMUXBANK２２BのMUX２４と双方向に通信する。そしてMUXBANK２２AのMUX２４が、ループバス２６Aを介して相互接続されている一方で、MUXBANK２２BのMUX２４は、ループバス２６Bを介して相互接続されている。そのため、各ディスクドライブ１８は、ループバス２６A及びループバス２６Bの両方を介してアクセス可能である。さらに、プロセッサブレイド１４Aがループバス２６Aと双方向に通信する一方で、プロセッサブレイド１４Bはループバス２６Bと双方向に通信する。プロセッサブレイド１４A及び１４Bは、直接相互接続され、ブレイドループ（ブレイド）バス３０を介して通信する。このため、プロセッサブレイド１４A及び１４Bは、対応するループバス２６を介して直接、または他方のプロセッサブレイド１４を介して間接的に、どのディスクドライブ１８とも双方向に通信できるとともに、相互に直接通信できる。

【００２０】最後に、記憶サブシステム１２について、本実施例のHANファイルサーバ１０においては、例えば、各ディスクドライブ１８は、簡単にユーザが置換できるキャリアに入れられたホットスワップファイバチャネルディスクドライブであり、ドライブ及びキャリアは、電気を供給し、MUXループバス２６A及び２６Bを含む中央平面にプラグ接続される。これにより、各デュアルポートドライブをMUX２４に、そしてMUX２４をループコントローラ２６と相互接続することができる。MUX２４はファイバチャネルMUXデバイスであり、ループコントローラ２６は、各MUXデバイスのパス選択を制御するマイクロコントローラを備え、各ディスクドライブ１８のデュアルポートのファイバチャネルMUXループバス２６A及び２６Bとの接続の実行又は解除を選択的に行う。MUX２４CA及び２４CB、MUX２４EA及び２４Eは同様に、ファイバチャネルMUXデバイスであり、記憶サブシステム１２をファイバチャネルループバスを介してコンピュートブレイド１４A及び１４BとEDISKA３２A及び３２Bとに接続する。コンピュートブレイドバス３０も同様にファイバチャネルバスである。

【００２１】b．制御／プロセッササブシステム１４（図１及び２）

前述のように、制御／プロセッササブシステム１４は、コンピュートブレイドバス３０を介して相互接続されるデュアルコンピュートブレイド（ブレイド）１４A及び１４Bからなる。コンピュートブレイド１４A及び１４Bは、共有記憶サブシステム１２の操作を制御する計算及び制御用のサブシステムを併せ持つ。コンピュートブレイド１４A及び１４Bは、HANファイルサーバ１０のクライアントにネットワークアクセスとファイルシステム機能とを独立して提供及びサポートし、相互バックアップと互いのネットワーク３４アクセス及びファイルシステム機能のためのサポートとを協調的に提供する。図１及び２に示すように、各ブレイド１４はネットワーク３４に接続された多数のネットワークポート（ポート）３４Pを備える。ネットワーク３４は、HANファイルサーバ１０とHANファイルサーバ１０のクライアント３４Cとの間の双方向データ通信接続を構成する。図に示すように、ネットワークには、例えば、クライアント３４Cに接続する複数のクライアントネットワーク３４Nと管理ネットワーク３４Mとが含まれ、さらにリモートクライアント３４Cに接続するルータ３４Rを含むこともできる。当業者には理解されるように、ネットワーク３４は、例えば、ローカルエリアネットワーク（LAN）、広域ネットワーク（WAN）、直接プロセッサ接続またはバス、ファイバオプティックリンク、あるいは前記の組み合わせから構成することができる。

【００２２】図２に示すように、各ブレイド１４は、メモリへの、そして通信コンポーネントのような他のエレメントへの緊密なアクセスを共有するデュアル処理ユニット３６A及び３６Bから構成される。各処理ユニット３６A及び３６Bは、フルオペレーティングシステムカーネルを実行する十分に機能的な計算処理ユニットであ

り、ファンクショナルマルチプロセッシング構造において協働する。例えば、後述されるような実装においては、一方の処理ユニット３６がＲＡＩＤ機能を実行し、他方の処理ユニット３６はネットワーク機能、プロトコルスタック機能、ＣＩＦＳ及びＮＳＦ機能、ファイルシステム機能を実行する。

【００２３】ｃ．ＨＡＮファイルサーバ１０の全体的なアーキテクチャ及びＨＡＮファイルサーバ１０の故障処理機構（図１及び２）

上述のように、本発明のＨＡＮファイルサーバ１０は階層及びピアドメインの集まり、すなわちノードあるいはサブシステムから構成され、各ドメインはファイルサーバの１つ以上のタスクまたは機能を実行するとともに故障処理機構を備えている。例えば、ＨＡＮファイルサーバ１０は、それぞれ、ネットワーク３４Ｎ、制御／プロセッササブシステム１４、記憶サブシステム１２を有する３つの階層ドメイン１０Ａ、１０Ｂ、１０Ｃから構成され、ファイルサーバの独立した及び相補的な機能を実行する。つまり、ドメイン１０Ａは、クライアント３４とＨＡＮファイルサーバ１０との間のクライアント／サーバ通信を提供し、ドメイン１０Ｂ、すなわち、制御／プロセッササブシステム１４は、ドメイン１０Ａのクライアント／サーバ通信をサポートするとともに高レベルファイルシステムトランザクションをサポートし、ドメイン１０Ｃ、すなわち、記憶サブシステム１２は、クライアントのファイルシステムをサポートする。制御／プロセッササブシステム１４は、２つのピアドメイン１０Ｄ及び１０Ｅ、すなわち、ブレイド１４Ａ及び１４Ｂからなり、並列機能、特にクライアント／サーバ通信機能及びより高い及び低いレベルのファイルシステム操作を実行し、それにより、クライアント通信及びファイル操作のタスクの負荷を分担する。後に詳細に説明されるように、ブレイド１４Ａ及び１４Ｂを備えたドメインはまた、クライアント／サーバ通信、ブレイド１４の相互通信、高レベルファイルシステム機能、記憶サブシステム１２で実行される低レベルファイルシステム機能の故障処理及びサポートを提供する独立して機能する故障処理機構を備える。各ブレイド１４は、処理ユニット３６Ａ及び３６Ｂに基づく２つの階層ドメイン１０Ｆ及び１０Ｇから構成されるドメインであり、ブレイド１４Ａ及び１４Ｂの機能を併せ持つ別個ではあるものの相補的な機能を実行する。後述するように、一方の処理ユニット３６は、高レベルファイル操作及びクライアント／サーバ通信を両機能のための故障処理機構に提供する上層ドメイン１０Ｆを形成する。他方の処理ユニット３６は、低レベルファイル操作及びブレイド１４の相互通信を提供する下層ドメイン１０Ｇを形成し、両機能及び上層ドメイン１０Ｆのサーバ機能と故障処理機構とをサポートする独立して機能する故障処理機構を備える。最後に、記憶サブシステム１２は、同様に、ディスクドライブ１８、すなわち、サーバの記憶エレメントを構成して、ブレイド１４のドメイン１０ＥによりサポートされるＲＡＩＤ機構を間接的にサポートする下層ドメイン１０Ｈと、ドメイン１０Ｄ及び１０Ｅとドメイン１０Ｈとの間の通信をサポートする記憶ループモジュール２０Ａ及び２０Ｂを備えたピア上層ドメイン１０Ｉ及び１０Ｊとから構成される。

【００２４】従って、以下に記述するように、各ＨＡＮファイルサーバ１０ドメインは、１つの中央統一機構あるいは調整機構なしに、互いに独立して別々に、しかしながら互いに協調的に動作する１つ以上の故障処理機構を直接あるいは間接的に有するまたは備える。そのため、あるドメインのコンポーネントの機能あるいは操作が故障しても、関連するドメインの対応するコンポーネントが後を引き継ぐ。さらに、以下に記述するように、ＨＡＮファイルサーバ１０の故障処理機構は、一箇所あるいは複数箇所に故障が起きても継続した機能を提供できるように、複数の異なる技術あるいは方法を外からはわからないように採用している。

【００２５】ＨＡＮファイルサーバ１０の全体構造及び操作をこれまで説明してきたが、以下には、ＨＡＮファイルサーバ１０の各ドメインをさらに詳細に、そしてＨＡＮファイルサーバ１０の故障処理機構の構造及び操作を説明する。

１．ブレイド１４の処理と制御コア

図２に、本実装のブレイド１４を示す。ブレイド１４は、デュアル処理ユニット３６Ａ及び３６Ｂの計算コアをそれぞれ形成するプロセッサ３８Ａ及び３８Ｂと、メモリコントローラハブ（ＭＣＨ）３８Ｃ、メモリ３８Ｄ、入出力コントローラハブ（ＩＣＨ）３８Ｅのような多数の共有エレメントとを備える。本実装において、例えば、プロセッサ３８Ａ及び３８Ｂは、それぞれ、内蔵のレベル２キャッシュを有するインテルペンティアムＩＩであり、ＭＣＨ３８Ｃ及びＩＣＨ３８Ｅはインテル８２０チップセットであり、メモリ３８ＤはＲＤＲＡＭあるいはＳＤＲＡＭの５１２ＭＢ以上からなる。

【００２６】図に示すように、プロセッサ３８Ａ及び３８Ｂは、パイプラインフロントサイドバス（ＦＳＢ）３８Ｆ及びＭＣＨ３８Ｃの対応するＦＳＢポート３８ＣＡを介してＭＣＨ３８Ｃと相互接続されている。当業者には理解されるように、ＭＣＨ３８Ｃ及びＭＣＨ３８ＣのＦＳＢポートは、プロセッサ３８Ａ及び３８Ｂからのメモリ参照の初期化及び受信と、プロセッサ３８Ａ及び３８Ｂからの入出力（Ｉ／Ｏ）及びメモリマップＩ／Ｏ要求の初期化及び受信と、メモリ３８Ｃからプロセッサ３８Ａ及び３８Ｂへのメモリデータの受け渡しと、メモリＩ／Ｏ要求から生じるメモリスヌープサイクルの初期化とをサポートする。さらに、ＭＣＨ３８Ｃはメモリ３８Ｄへのメモリポート３８ＣＢと、ＩＣＨ３８Ｅへのハブリンクバス３８Ｇに接続するハブリンクポート３８ＣＣ

と、業界標準パーソナルコンピュータ相互接続（ＰＣＩ）バスとして機能する４つのＡＧＰポート３８ＣＤとを備えている。各ＰＣＩバスは、インテル２１１５４チップのようなプロセッサツープロセッサブリッジユニット（Ｐ－Ｐブリッジ）３８Ｈへのプロセッサに接続されている。

【００２７】ＩＣＨ３８Ｅは、ＭＣＨ３８Ｃへのハブリングバス３８Ｇに接続するハブリンクポート３８ＥＡ、ファームウェアメモリ３８Ｉに接続するファームウェアポート３８ＥＢ、ハードウェアモニタ（ＨＭ）３８Ｊに接続するモニタポート３８ＥＣ、ブートドライブ３８Ｋに接続するＩＤＥドライブポート３８ＥＤ、スーパーＩ／Ｏデバイス（スーパーＩ／Ｏ）３８Ｌに接続するＩ／Ｏポート３８ＥＥ、他のエレメントと共に、ＶＧＡデバイス（ＶＧＡ）３８Ｍ及び管理ローカルエリアネットワークデバイス（ＬＡＮ）３８Ｎに接続するＰＣＩポート３８ＥＦを含んでいる。当業者には上記の説明で十分理解されるであろう。

【００２８】２．ブレイド１４のパーソナルコンピュータ互換サブシステム

ＩＣＨ３８Ｅ、スーパーＩ／Ｏ３８Ｌ、ＶＧＡ３８Ｍは併せてパーソナルコンピュータ（ＰＣ）互換サブシステムを構成し、ローカル制御及び表示の目的でＨＡＮファイルサーバ１０のためのＰＣ機能及びサービスを提供する。この目的のために、当業者には理解されるように、ＩＣＨ３８Ｅは、ＩＤＥコントローラ機能、ＩＯ　ＡＰＩＣ、８２Ｃ５９ベースのタイマ及びリアルタイムクロックを備える。スーパーＩ／Ｏ３８Ｌは、例えば、標準マイクロシステムデバイスＬＰＣ４７Ｂ２７ｘであってもよく、８０４２キーボード／マウスコントローラ、２．８８ＭＢスーパーＩ／Ｏフロッピィディスクコントローラ、フル機能デュアルシリアルポートを提供する。一方、ＶＧＡ３８Ｍは、例えば、１ＭＢフレームバッファメモリをサポートするシーラスロジック（Ｃｉｒｒｕｓ　Ｌｏｇｉｃ）６４ビットビジュアルメディア（ＶｉｓｕａｌＭｅｄｉａＲ）アクセラレータＣＬ－ＧＤ５４４６－ＱＣであってもよい。

【００２９】３．ブレイド１４のファームウェア及びＢＩＯＳサブシステム

ＩＣＨ３８Ｅ及びファームウェアメモリ３８Ｉは、併せて、通常のファームウェア及びＢＩＯＳ機能を実行するファームウェア及びＢＩＯＳサブシステムを構成し、その機能には、ブレイド１４Ａ及び１４Ｂリソースのパワーオンセルフテスト（ＰＯＳＴ）及びフル設定が含まれる。例えば、ＡＭＩ／Ｐｈｏｅｎｉｘから利用できるような標準ＢＩＯＳであるファームウェア及びＢＩＯＳは、１ＭＢのフラッシュメモリを備えたファームウェアメモリ３８Ｉに存在する。ＰＯＳＴが完了すると、ＢＩＯＳは上述したＰＣＩバスをスキャンし、このスキャンの間、上述及び後述する２つのＰＣＩツーＰＣＩブリッジを設定し、以下に記述するバックエンド及びフロントエンドＰＣＩバス上のファイバチャネル及びＬＡＮコントローラの存在を検出してＰＣＩアドレス空間にマップする。この情報は、ＰＣ互換ＩＯやメモリサイズなどの他の標準サイジング情報とともにＩＯサブシステムのトポロジを記述するＭＰ対応テーブルに書き留められ、ＰＯＳＴは単純なパスチェックとメモリ診断とを実行する。ＰＯＳＴの終了後、綿密なリブート診断パッケージを含むフラッシュレジデントユーザバイナリコードセグメントがロードされる。リブート診断パッケージはまた、ファイバチャネルデバイスを初期化するとともに、パターンセンシティブデータを使ってデータパス及びＤＲＡＭセルを試験することにより、コンピュートブレイド上のコンポーネントの整合性をチェックする。診断が実行されると、制御はＢＩＯＳあるいはブートストラップユーティリティに戻される。制御がＢＩＯＳに移される場合は、システムはブートを続け、制御がブートストラップユーティリティに移される場合は、ブートブロックがファイバディスクから読み取られ、制御は新しくロードされたオペレーティングシステムのイメージに引き渡される。さらに、このサブシステムは、全体のシステム管理アーキテクチャをサポートする、エラーチェックロジック、環境モニタリング、エラー及びスレッショルドロギングなどの機能を提供する。最下層レベルでは、内蔵プロセッサキャッシュパリティ／ＥＣＣエラー、ＰＣＩバスパリティエラー、ＲＤＲＡＭ　ＥＣＣエラー、フロントサイドバスＥＣＣエラーを含むハードウェアエラー及び環境スレッショルドチェックが実行される。エラー及び超過の環境スレッショルドイベントは、ＤＭＩ互換レコードフォーマットでフラッシュプロムの一部にロギングされる。

【００３０】４．ブレイド１４のＩ／Ｏバスサブシステム

最後に、ＭＣＨ３８Ｃ及びＩＣＨ３８Ｅは、ブレイド１４の２つの入出力（Ｉ／Ｏ）バスサブシステムをサポートする。うち一方はＭＣＨ３８Ｃによってサポートされるバックエンドバスサブシステム（ＢＥ　ＢｕｓＳｙｓ）３８Ｏであり、前述のブレイド１４及び記憶サブシステム１２の対応するループパス２６間の双方向接続と、コンピュートブレイドバス３０を介したブレイド１４Ａ及び１４Ｂ間の双方向接続とを提供する。他方はＩＣＨ３８Ｅによってサポートされるフロントエンドバスサブシステム（ＦＥ　ＢｕｓＳｙｓ）３８Ｐであり、前述のネットワーク３４への及びそれからの双方向接続を提供する。ネットワーク３４は、前述のように、例えば、ローカルエリアネットワーク（ＬＡＮ）、広域ネットワーク（ＷＡＮ）、直接プロセッサ接続またはバス、ファイバオプティックリンク、あるいは前記の組み合わせであることができる。

【００３１】まず、ＢＥ　ＢｕｓＳｙｓ３８Ｏについて

考えると、上述のように、ＭＣＨ３８Ｃは、業界標準パーソナルコンピュータ相互接続（ＰＣＩ）バスとして機能する４つのＡＧＰポート３８ＣＤをサポートする。各ＡＧＰポート３８ＣＤは、インテル２１１５４チップのようなプロセッサツープロセッサブリッジユニット（Ｐ－Ｐブリッジ）３８Ｈに接続される。Ｐ－Ｐブリッジ３８Ｈは、例えば、タックライト（Ｔａｃｈ　Ｌｉｔｅ）ファイバチャネルコントローラから構成される２つのファイバチャネルコントローラ（ＦＣＣ）３８Ｑの双方向バスポートに接続される。ＦＣＣ３８Ｑの並列ファイバチャネルインターフェイスは、２つの対応するシリアライザ／デシリアライザデバイス（ＳＥＲ－ＤＥＳ）３８Ｒの並列ファイバチャネルインターフェイスに接続されている。一方のＳＥＲ－ＤＥＳ３８Ｒのシリアルインターフェイスはコンピュートブレイドバス３０に接続され、他方のデュアルブレイド１４への通信接続を提供する。他方のＳＥＲ－ＤＥＳ３８Ｒのシリアルインターフェイスは記憶サブシステム１２の対応するループバス２６に接続されている。

【００３２】ＦＥ　ＢｕｓＳｙｓ３８Ｐでは、上述のように、ＩＣＨ３８ＥがＰＣＩポート３８ＥＦを備えており、図に示すように、ＰＣＩポート３８ＥＦは、ＰＣＩバスツーＰＣＩバスブリッジユニット（Ｐ－Ｐブリッジ）３８Ｈと双方向に接続されている。Ｐ－Ｐブリッジ３８Ｈは、例えば、双方向３２ビット３３ＭＨｚフロントエンドＰＣＩバスセグメントをサポートするインテル２１１５２から構成される。フロントエンドＰＣＩバスセグメントは、ネットワーク３４に接続する１群の双方向ネットワークデバイス（ＮＥＴＤＥＶ）３８Ｔに接続されていて、ＮＥＴＤＥＶ３８Ｔは、例えば、インテル８２５５９　１０／１００イーサネットコントローラデバイスである。前述のように、ネットワーク３４は、例えば、ローカルエリアネットワーク（ＬＡＮ）、広域ネットワーク（ＷＡＮ）、直接プロセッサ接続またはバス、ファイバオプティックリンク、あるいは前記の組み合わせであることができ、ＮＥＴＤＥＶ３８Ｔはそれに応じて選択されることが理解されるであろう。

【００３３】最後に、ＢＥ　ＢｕｓＳｙｓ３８Ｏ及びＦＥ　ＢｕｓＳｙｓ３８Ｐについて、本実施例においては、ＢＥ　ＢｕｓＳｙｓ３８Ｏ及びＦＥ　ＢｕｓＳｙｓ３８Ｐの両方がＰＣＩタイプのバスであり、そのため、共通の割り込み構造を有している。このため、ＢＥ　ＢｕｓＳｙｓ３８Ｏ及びＦＥ　ＢｕｓＳｙｓ３８ＰのＰＣＩ割り込みは、ＢＥ　ＢｕｓＳｙｓ３８ＯのＰＣＩバスデバイスがＦＥ　ＢｕｓＳｙｓ３８ＰのＰＣＩバスデバイスと割り込みを共有しないようにルーティングされる。

【００３４】ｃ．ＨＡＮファイルサーバ１０の操作（図１、２、３）

１．ＨＡＮファイルシステム１０の全体的な操作

上述のように、ＨＡＮファイルシステム１０は、デュアルコンピュートブレイド１４を備え、各コンピュートブレイド１４は記憶サブシステム１２の全てのディスクドライブ１８への完全なアクセスと、全てのクライアントネットワーク３４Ｎへの接続とを有し、それぞれ独立してＨＡＮファイルシステム１０の全ての機能及び操作を実行できる。ブレイド１４の機能及び操作構造の概略図を図３に示す。図３は、ブレイド１４Ａ及び１４Ｂのうちの一方を示し、他方のブレイド１４は図のブレイド１４と同一であり、かつミラーイメージであることが理解されるだろう。

【００３５】ブレイド１４の内部では、上述のように、デュアル処理ユニット３６Ａ及び３６Ｂが、例えば、メモリコントローラハブ（ＭＣＨ）３８Ｃ、メモリ３８Ｄ、入出力コントローラハブ（ＩＣＨ）３８Ｅのような、多数のブレイド１４エレメントを共有している。処理ユニット３６Ａ及び３６Ｂはそれぞれ、互いに独立しながらも協調的に動作し、それぞれがメモリ３８Ａに存在するリアルタイムオペレーティングシステム（ＯＳ）４０の別々のコピーを実行する。ＯＳ４０の各コピーは、例えば、処理ユニット３６Ａ及び３６Ｂの対応する一方のために、基本メモリ管理、タスクスケジューリング、同期機能、他の基本オペレーティングシステム機能を提供する。処理ユニット３６Ａ及び３６Ｂは、共有メモリ３８Ａに設けられたメッセージパッシング機構（メッセージ）４２を介して通信し、メッセージは、例えば、Ｉ／Ｏの開始、Ｉ／Ｏの終了、ディスク故障のようなイベント通知、ステータスクエリー、ブレイドバス３０を介してミラーリングされる、ファイルシステムジャーナルのような重要なデータ構造のミラーリングのために規定される。初期設定時、各ブレイド１４はＯＳ４０と、ＲＡＩＤファイルシステム及びネットワークイメージとの両方のコピーをバックエンドディスクドライブ１８からロードする。それぞれ処理ユニット３６Ａ及び３６Ｂの一方を実行する２つのＲＡＩＤカーネルは、その後、ＯＳ４０の２つのインスタンス間でブレイド１４のメモリ３８Ａを協力して分割し、ＯＳ４０カーネルのコピーがロードされた後、処理ユニット３６Ａ及び３６Ｂの操作を開始する。初期設定の後、ＯＳ４０カーネルはメッセージ４２を介して通信する。

【００３６】図３に示すように、各ブレイド１４の内部で、処理ユニット３６Ａ及び３６Ｂの一方はバックエンドプロセッサ（ＢＥＰ）４４Ｂと称されて動作する。そして、上述のように、ＲＡＩＤ設定ディスクへの及びそれからのデータの書き込み及び読み出しのためのブロック記憶システムとして動作するとともに、ＲＡＩＤ機構（ＲＡＩＤ）４６を備える。ＲＡＩＤ４６には、ＲＡＩＤデータ記憶及びバックアップ機能を実行するＲＡＩＤファイル機構（ＲＡＩＤＦ）４６Ｆと、ＲＡＩＤ関連のシステムモニタリング機能及び以下に示す他の機能を実

行するＲＡＩＤモニタ機構（ＲＡＩＤＭ）４６Ｍとが含まれる。処理ユニット３６Ａ及び３６Ｂの他方はフロントエンドプロセッサ（ＦＥＰ）４４Ｆと称されて動作し、クライアントとディスクレジデントブロック記憶システムとの間でデータを移動するための全てのネットワーク及びファイルシステム操作、そして、ネットワークドライバ、ＣＩＦＳ及びＮＦＳプロトコルを含むプロトコルスタックのサポートとジャーナルファイルシステムの維持とを含めたＢＥＰ４４Ｂの対応するＲＡＩＤ機能を実行する。

【００３７】ブロック記憶システム操作に加えて、ＢＥＰ４４Ｂの機能には、ＲＡＩＤＦ４６Ｆ及びＲＡＩＤＭ４６Ｍを介してのコアＲＡＩＤファイルシステムサポートアルゴリズムの実行、ディスクドライブ１８の操作のモニタリング、自身が存在するブレイド１４及びピアブレイド１４の両方の操作及び状態のモニタリング、管理機能への故障の連絡が含まれる。図２及びＢＥ　ＢｕｓＳｙｓ３８Ｏについて上述したように、ＢＥＰ４４Ｂはまた、ＢＥ　ＢｕｓＳｙｓ３８Ｏとブレイドバス３０とを介してブレイド１４Ａ及び１４Ｂ間の通信を、そしてＢＥ　ＢｕｓＳｙｓ３８Ｏと記憶サブシステム１２の対応するループパス２６とを介してディスクドライブ１８との通信をサポートする。ＲＡＩＤＭ４６Ｍはまた、ブレイド１４の電源装置をモニタし、停電の際には適切な処理を実行する。例えば、ディスクドライブ１８に重要なデータ構造の緊急書き込みを行ったり、処理ユニット３６Ａ及び３６Ｂの生き残った方が適切な処理を開始できるように処理ユニット３６Ａ及び３６Ｂの一方に通知をする。ＢＥＰ４４Ｂはさらに、確実なブートストラップサポート機能を提供し、それによりランタイムカーネルがディスクドライブ１８に保存され、システムブートの際ロードされることができる。

【００３８】ＦＥＰ４４Ｆは、ブレイド１４の全てのネットワーク３４関連機能及び操作を実行するネットワーク機構（ネットワーク）４８を備え、ＦＥ　ＢｕｓＳｙｓ３８Ｐ及びＮｅｔＤｅｖ３８Ｔのエレメントを含んでいる。例えば、ネットワーク４８は、ＦＥ　ＢｕｓＳｙｓ３８Ｐを含むネットワーククライアントに利用可能なリソースを管理及び提供し、ネットワーク３４を介してクライアント３４ＣにＨＡＮファイルシステム１０へのアクセスを提供する。後述するように、ネットワーク４８はまた、ＦＥＰ４４Ｆに存在する通信フェイルオーバー機構と、ここに記載されるその他の高可用性機能とをサポートする。

【００３９】ＦＥＰ４４Ｆはまた、ジャーナルファイルシステム（ＪＦｉｌｅ）５０を含む。ジャーナルファイルシステム（ＪＦｉｌｅ）５０は、ネットワーク４８を介してＨＡＮファイルシステム１０のクライアントと、そしてメッセージ４２を介してＲＡＩＤＭ４６ＦのＲＡＩＤファイルシステム機能と通信する。図に示すように、ＪＦｉｌｅ５０は、ＪＦｉｌｅ５０のファイルシステム機能を実行するファイルシステム機構（ＦＳＭ）５０Ｆと、ＦＳＭ５０Ｆと相互作用してそれぞれデータトランザクションのデータ及び操作をキャッシュし、データトランザクションのジャーナルを維持する内蔵書き込みキャッシュ（ＷＣａｃｈｅ）５０Ｃ及びトランザクションログ（ログ）５０Ｌとを含む。ログ５０Ｌには、要求されたデータトランザクションを表すログエントリ（ＳＥ）５０Ｅを生成するためのログジェネレータ（ＬＧｅｎ）５０Ｇと、ＳＥ５０Ｅを記憶するログメモリ（ＬｏｇＭ）５０Ｍとが含まれる。ＬｏｇＭ５０Ｍの大きさは、以下に記述されるように、ジャーナルされるべきデータトランザクションの数に依存する。図に示すように、ＢＥＰ４４Ｂには、ＷＣａｃｈｅ５０Ｃと通信して、ＷＣａｃｈｅ５０Ｃの中身をミラーリングするキャッシュミラー機構（ＣＭｉｒｒｏｒ）５４Ｍが含まれる。さらに、各ブレイド１４のログ５０Ｌは、反対側のピアブレイド１４に存在するログ５０Ｌのミラー機構（ＬＭｉｒｒｏｒ）５４Ｌによってミラーリングされ、各ブレイド１４のログ５０Ｌは、メッセージ４２、ＢＥ　ＢｕｓＳｙｓ３８Ｏ、ブレイドバス３０を含むパスを介して対応するＬＭｉｒｒｏｒ５４Ｌと通信する。

【００４０】最後に、ＦＥＰ４４Ｆには、ステータスモニタ機構（モニタ）５２が含まれる。モニタ５２は、ＨＡＮファイルシステム１０の変更に関するＢＥＰ４４Ｂからの通知をモニタし、その変更を受けて適切な処理を開始する。この通知には、例えば、ＲＡＩＤグループに新しく挿入されたディスクのバインディングに関する、あるいは故障したディスクのためのＳＮＭＰトラップを起動するＲＡＩＤＭ４６Ｍからの通知が含まれ、モニタ５２により開始される操作には、例えば、以下に記述するように、ＲＡＩＤ機能が非常に重大なエラーに遭遇した場合等に、ＨＡＮファイルサーバ１０の故障処理機構によりフェイルオーバー動作を開始すること、あるいはブレイド１４を完全にシャットダウンすることが含まれる。

【００４１】２．ＨＡＮファイルサーバ１０のファイルシステム機構の操作（図１、２、３）

上記及び図３に示したように、ＨＡＮファイルサーバ１０のファイルサーバ機構は、３つの主要なコンポーネントあるいは層を含む。１つ目の最上層は、ブレイド１４Ａ及び１４Ｂそれぞれのフロントエンドプロセッサ４４Ｆに存在するＷＣａｃｈｅ５０Ｃ及びＬｏｇ５０Ｌを含むＪＦｉｌｅ５０のファイルシステム機構である。最下層には、ディスクドライブ１８を備えた記憶サブシステム１２と、ブレイド１４Ａ及び１４Ｂそれぞれの ＢＥＰ４４Ｂに存在するブロック記憶システム機能及びＲＡＩＤＦ４６Ｆ機能とが含まれる。ＨＡＮファイルサーバ１０ファイルシステム機構の３番目の層あるいはコンポーネントは、ファイルシステム機構の操作に影響する故障

を検出して処理し、ファイルシステム故障からの回復を行う故障処理機構から構成される。上層及び下層ファイルシステムエレメントの構造及び操作はすでに上述されており既知のエレメントと類似しているので当業者にはよく理解されることであろう。そのため、本実施例のＨＡＮファイルサーバ１０ファイル機構のこれらのエレメントは、本発明を完全に理解するのに必要でない限りここでは詳細に説明されない。以下の記述は、その代わりに、ＨＡＮファイルサーバ１０ファイル機構の故障処理機構、特にＨＡＮファイルサーバ１０の上層レベルのファイルシステムエレメントの操作に関する故障処理機構に焦点をあてる。

【００４２】上述のように、ＨＡＮファイルサーバ１０ファイル機構の第３のコンポーネントは、ＨＡＮファイルサーバ１０コンポーネントの損失から生じるデータの損失に対する保護を提供するミラーリング機構から構成される。図３に示すように、ミラーリング機構には、各ブレイド１４毎に、ブレイド１４のＢＥＰ４４Ｂに存在するキャッシュミラー機構（ＣＭｉｒｒｏｒ）５４Ｍと、反対側のピアブレイド１４のＢＥＰ４４Ｂに存在するログミラー機構（ＬＭｉｒｒｏｒ）５４Ｌとが含まれる。ＣＭｉｒｒｏｒ５４Ｍは、メッセージ４２を介してＪＦｉｌｅ５０のＷＣａｃｈｅ５０Ｃと通信する継続動作キャッシュミラーリング機構である。ログ５０Ｌは、ピアブレイド１４のＢＥＰ４４Ｂに存在するＬＭｉｒｒｏｒ５４Ｌにより要求に応じてミラーリングされ、メッセージ４２、ＢＥ　ＢｕｓＳｙｓ３８Ｏ、コンピュートブレイドバス３０を介して対応するＬｏｇＭ５０Ｍと通信する。これにより、クライアントに通知される前に、ブレイド１４Ａあるいは１４Ｂの一方を介したファイルシステムへの全データ変更が、ブレイド１４Ａあるいは１４Ｂの他方に反映される。これに関連して、本実施例においては、ログ５０Ｌのミラーリングは、各ファイルシステムトランザクションの処理中に実行される。そのため、トランザクションログミラーリングのレイテンシは実際のファイルシステムトランザクションの実行により限度ぎりぎりまで掩蔽される。最後に、ＲＡＩＤＦ４６Ｆによりサポートされ提供されるディスクドライブ１８ファイルシステム、制御、モニタリング、データ回復／再構築機能は、ＨＡＮファイルサーバ１０データ保護機構の一部でもあり、記憶サブシステム１２内部へのデータミラーリング法を使用していることが理解されるだろう。

【００４３】以下に記述されるように、これらのミラーリング機構は、よって、故障のタイプによって、ブレイド１４における故障を処理する数多くの代替法をサポートしている。例えば、ブレイド１４の一方が故障した際、生き残ったブレイド１４は、そのＬＭｉｒｒｏｒ５４Ｌに保存されたファイルトランザクションを読み取り、故障したブレイド１４が復帰したときに故障していたブレイド１４に戻す。その際には、復帰したブレイド１４により失われたファイルトランザクションが再実行され回復される。他の方法では、ブレイド１４のネットワーク３４フェイルオーバー機構について以下に記述するように、故障したブレイド１４あてのファイルトランザクションが、ブレイド１４間のブレイドバス３０のパスを介して、あるいはブレイド１４のネットワーク３４ファイルオーバー機構によって生き残ったブレイド１４へのクライアントのリダイレクションにより、生き残っているブレイド１４にリダイレクトされる。生き残ったブレイド１４は、それにより、故障したブレイド１４あてのファイルトランザクションの実行を引き継ぐ。以下に記述するように、生き残ったブレイド１４は、この操作の一部として、そのＬＭｉｒｒｏｒ５４Ｌに保存されている故障したブレイド１４からのファイルトランザクションを再実行することにより故障したブレイド１４の失われたファイルトランザクションを再実行して回復するか、あるいは、故障したブレイド１４が復帰した後に故障していたブレイド１４にファイルトランザクションを読み戻す。これにより、故障の際の故障したブレイド１４上のファイルシステムの状態が再構築され、確認済みのトランザクションのために、故障したブレイドからデータが失われることはない。

【００４４】３．ＨＡＮファイルサーバ１０の通信機構の操作（図１、２、３）

図１、２、３に示すように、本発明に組み込まれているＨＡＮファイルサーバ１０の通信機構は、３つのレベルあるいは層の通信機構から構成されるとみなすことができる。説明のために、最上層レベルは、クライアント３４Ｃと、ＨＡＮファイルサーバ１０によってサポートされるクライアントファイルシステム構造との間のファイルトランザクション通信のためのネットワーク３４関連通信機構、及び、関連する通信故障処理機構から構成される。通信機構の中間層には、ブレイドバス３０及びメッセージ４２を介したブレイド１４Ａ及び１４Ｂ間の通信をサポートする通信機構と、関連する通信故障処理機構とが含まれる。通信機構の最下層には、ブレイド１４及び記憶サブシステム１２間、そして記憶サブシステム１２のエレメント間の通信パス及び機構とが含まれる。前記は、すでに説明されており、本発明を理解するために必要でない限りさらには説明されない。

【００４５】まず、ＨＡＮファイルサーバ１０の通信機構の上層レベルについて考える。図３に示すように、ブレイド１４Ａ及び１４ＢそれぞれのＦＥＰ４４Ｆに存在するネットワーク機構（ネットワーク）４８は、ＴＣＰ／ＩＰプロトコルスタック（ＴＣＰ／ＩＰスタック）５８を含むネットワークスタックオペレーティングシステム（ＮｅｔＳＯＳ）５６とネットワークデバイスドライバ（ＮｅｔＤＤ）６０とを含み、以下に記述するように、これらの機構には、単一ポート３４Ｐの故障、ネッ

トワーク３４の故障、ブレイド１４全体の故障を調整して処理する機能が含まれる。これに関連して、本文の他の箇所にも記載するように、ネットワーク３４は、例えば、ローカルエリアネットワーク（ＬＡＮ）、広域ネットワーク（ＷＡＮ）、直接プロセッサ接続またはバス、ファイバオプティックリンク、あるいは前記の組み合わせから構成されることができ、ＮＥＴＤＥＶ３８Ｔ及びＮｅｔＤＤ６０はそれに応じて実装される。

【００４６】また、図３に示され、ＨＡＮファイルサーバ１０の通信機構の高可用性について以下に説明されるように、各ネットワーク４８はさらに、クライアントルーティングテーブル（ＣＲＴ）４８Ａを含む。ＣＲＴ４８Ａは、ブレイド１４によりサポートされるクライアント３４Ｃに付随するルーティング及びアドレス情報を含むクライアントルーティングエントリ（ＣＲＥ）４８Ｅと、反対側のピアブレイド１４によってサポートされるクライアント３４ＣのＣＲＥ４８Ｅとを保存する。当業者には理解されるように、ＣＲＥ４８Ｅは、ネットワーク４８によって、所定のクライアント３４Ｃへファイルトランザクション通信を送るために利用されることができ、必要であるならば、ブレイド１４に割り当てられたクライアント３４Ｃから受領したファイルトランザクション通信を識別、あるいは確認するために利用されることもできる。図に示すように、各ネットワーク４８にはまた、ブレイドルーティングテーブル（ＢＲＴ）４８Ｂが含まれる。ＢＲＴ４８Ｂは、ブレイド１４にアクセス可能でブレイド１４によって共有されるネットワーク３４通信パスに関するアドレス及びルーティング情報を含み、これにより、ブレイド１４間の利用可能な通信パスを形成する。典型的な本実装のネットワーク４８において、ＣＲＴ４８Ａ及びＢＲＴ４８Ｂ情報は、ブレイドバス３０を含む通信パスを介してブレイド１４Ａ及び１４Ｂ間で通信されるが、例えば、ネットワーク３４Ｍを介して各ブレイド１４に提供されることもできる。

【００４７】ＨＡＮファイルサーバ１０のネットワーク３４通信機構の全体的な操作を説明する。図１及び２を見ると、ＨＡＮファイルサーバ１０の各ブレイド１４は、ネットワーク３４と接続して通信する複数のポート３４Ｐをサポートしている。例えば、本実装において、各ブレイド１４は合計５つのポート３４Ｐをサポートしていて、うち４つのポート３４Ｐはネットワーク３４Ｎに接続されてクライアント３４Ｃにサービスを提供し、１つのポートは、ＨＡＮファイルサーバ１０の管理のために予約されて管理ネットワーク３４Ｍに接続されている。図に示すように、ブレイド１４Ａ及び１４Ｂそれぞれの対応するポート３４Ｐは同じネットワーク３４に接続されており、そのため、各ネットワーク３４は、対応するポート３４Ｐを介して、ブレイド１４Ａ及び１４Ｂそれぞれに接続される。本実施例において、ＨＡＮファイルサーバ１０のポート３４Ｐは、１０個の異なるＩＰアドレス、すなわち、各ポートにつき１アドレスを設定され、ブレイド１４のそれぞれ対応する組み合わせのポート３４Ｐのポート３４Ｐが同じネットワーク３４に接続されている。そのため、各ネットワーク３４は、２つのアドレス、すなわちブレイド１４Ａ及び１４Ｂそれぞれの一方へのアドレスを介してＨＡＮファイルサーバ１０をアドレス指定することができる。ＨＡＮファイルサーバ１０の各クライアントが割り当てられるポート３４Ｐは、従来技術であり当業者には簡単に理解されるように、クライアントに存在するＡＲＰテーブルにより各クライアント内で決定される。さらに、図２に示すように、クライアント３４Ｃは、ＨＡＮファイルサーバ１０がデフォルトのルートを設定されるかまたはＲＩＰまたはＯＳＰのようなルーティングプロトコルを備える場合、直接接続されたネットワーク３４通信のうちの一方を介して、あるいは任意のルータ３４Ｒを介して、ＨＡＮファイルサーバ１０にアクセスできる。ＨＡＮファイルサーバ１０の別の実装では、各クライアント３４Ｃは、複数のネットワーク３４を介してＨＡＮファイルサーバ１０のポート３４Ｐに接続されることができ、ネットワーク３４は、以下に記述するように、クライアント３４ＣのＡＲＰテーブル及びＨＡＮファイルサーバ１０を適切に改良することにより、ローカルエリアネットワーク（ＬＡＮ）、広域ネットワーク（ＷＡＮ）、直接プロセッサ接続またはバス、ファイバオプティックリンク、あるいは前記の組み合わせのような異なる技術を利用することができる。

【００４８】図３に示すように、ブレイド１４Ａ及び１４Ｂそれぞれの各ＦＥＰ４４Ｆに存在するネットワーク４８機構はさらに、ＣＩＦＳ６２及びＮＦＳ６４ネットワークファイルシステムと、その他の必要なサービスとを備える。図３には示されていないこれらの付加的なサービスには、以下のものが含まれる。

【００４９】ＮＥＴＢＩＯＳ　－　リモートリソースにアクセスするためにＰＣクライアントによって使用されるマイクロソフト／ＩＢＭ／インテルプロトコル。このプロトコルの重要な特徴の１つは、サーバ名をトランスポートアドレスに変更することであり、サーバは、共有資源、すなわち、¥¥server¥shareを識別するためにクライアントにより用いられるＵＮＣ名のコンポーネントとなる。ＨＡＮファイルサーバ１０では、サーバはブレイド１４Ａまたは１４Ｂを表す。ＮＥＴＢＩＯＳはまた、ＣＩＦＳ６２パケットフレーミングを提供し、ＨＡＮファイルサーバ１０はＲＦＣ１００１及びＲＦＣ１００２に規定されるようなＴＣＰ／ＩＰに優先してＮＥＴＢＩＯＳを使用する。

【００５０】ＳＮＭＰ　－　Ｓｉｍｐｌｅ　Ｎｅｔｗｏｒｋ　Ｍａｎａｇｅｍｅｎｔ　Ｐｒｏｔｏｃｏｌ。ＨＡＮファイルサーバ１０に、エージェントと呼ばれる処理を提供する。エージェントは、システムについての情報

を提供するとともに、通常でないイベントが起きた際、トラップを送信する機能を提供する。

【００５１】ＳＭＴＰ　－　Ｓｉｍｐｌｅ　Ｍａｉｌ　Ｔｒａｎｓｐｏｒｔ　Ｐｒｏｔｏｃｏｌ。通常でないイベントが起きた際、電子メールメッセージを送信するためにＨＡＮファイルサーバ１０により用いられる。

ＮＦＳ　－　サンマイクロシステムズネットワーク情報サービス。ＮＳＦファイルシステムへのアクセス制御に用いられるユーザＩＤを識別するためにＮＦＳサーバによって用いられるプロトコルを提供する。

【００５２】ＲＩＰ　－　動的ルーティングプロトコル。ルータ３４Ｒのようなルータの背後で動作しているクライアントのサポートによりネットワークトポロジを明らかにするために使用される。本実装のＨＡＮファイルサーバ１０においては、このプロトコルは、ルーティング情報のモニタのために受動モードで動作する。別の実装においては、ユーザがシステム初期設定の間にデフォルトルートを設定または明示してもよい。

【００５３】本発明の説明では、ＨＡＮファイルサーバ１０の正常動作時は、各ネットワーク４８のエレメント、すなわち、ＮｅｔＳＯＳ５６、ＴＣＰ／ＩＰスタック５８、ＮｅｔＤＤ６０、ＣＲＴ４８Ａは、クライアント３４ＣとＨＡＮファイルサーバ１０との間のネットワーク通信操作を実行するのに当業者には明らかな従来方法で動作することが当業者には理解されるであろう。このため、ＨＡＮファイルサーバ１０のこれらの機能についてはこれ以上説明をしない。以下はＨＡＮファイルサーバ１０のネットワーク関連通信機構の高可用性に焦点をあてて説明する。

【００５４】４．ＨＡＮファイルサーバ１０の通信故障処理機構(図１、２、３)

ａ．ネットワーク通信故障機構

通信または接続故障が簡単に検出される一方、どのコンポーネントが故障したのかを見極め、どんな訂正手段をとるのが適当かを判断することが難しくかつ複雑であることは当業者には明白に理解されることであろう。例えば、故障の可能性のあるソースには、ポート３４Ｐ、あるいはポート３４Ｐとネットワーク３４のハブまたはスイッチとの間のリンク、あるいはブレイド１４間のネットワークのパーティションが含まれるがこれに限定されるわけではない。しかしながら、ＨＡＮファイルサーバ１０は、ブレイド１４故障と同様に、１つ以上のネットワーク３４インターフェイス故障及び、異なるタイプのネットワーク３４故障とに対処できるＩＰネットワーク通信サービスを提供し、さらに、さまざまな故障を徐々に減少させる機能をサーバシステムに提供するために、異なるクラスあるいはタイプの故障を処理する多数の協調的あるいは補足的な機構を実装する。例えば、ブレイド１４のポート３４Ｐインターフェイス故障の際、ＨＡＮファイルサーバ１０は、ブレイド１４Ａ及び１４Ｂ間のコンピュートブレイドパス３０接続を利用して、ネットワークトラフィックをピアブレイド１４上の機能している対応ポート３４Ｐからポート３４Ｐが故障したブレイド１４へ転送することができる。この機能により、１つのネットワークポート３４Ｐの故障によりブレイド１４全体が動かなくなるのが防がれ、その結果、ブレイド１４によってサポートされるファイルシステムを移動する必要がなくなる。この機能はまた、故障が異なるネットワーク３４上で起きる限り、すなわち、故障がブレイド１４上の対応するポート３４Ｐの両方に起きない限り、片方あるいは両方のブレイド１４上での複数のネットワークポート３４Ｐ故障を調整できることが明らかである。各ネットワーク３４のブレイド１４の一方で少なくとも１つのポート３４Ｐが機能する限り、クライアントには故障が起きていることがわからない。

【００５５】ＨＡＮファイルサーバ１０の高可用性通信機構は、各ブレイド１４ドメインに存在する通信フェイルオーバー機構（ＣＦａｉｌ）６６により提供される。ＣＦａｉｌ６６は、各ブレイド１４のネットワーク４８の機構とブレイド１４Ａ及び１４Ｂのメッセージ４２機構とについての通信故障処理のために別々に動作するものの協調的な機構を含む。

【００５６】まず、ネットワーク４８、すなわち、クライアント３４Ｃ及び制御／プロセッササブシステム１４ドメイン間の通信についてのＣＦａｉｌ６６の機能及び操作について考える。ＣＦａｉｌ６６はＩＰパススルーと呼ばれる操作を実行し、これにより、一方のブレイド１４に関連する故障したネットワーク３４サービスは、反対側のピアブレイド１４の故障していない対応ポート３４Ｐに移され、以下に記述するように、ブレイド１４を通る代わりのパスを介してルーティングされる。図３に示すように、各ＣＦａｉｌ６６には、ブレイド１４のＦＥＰ４４Ｆに存在する通信モニタリング処理／プロトコル機構（ＣＭｏｎｉｔｏｒ）６６Ｃが含まれる。ＣＭｏｎｉｔｏｒ６６Ｃは、ブレイド１４Ａ及び１４ＢのＮｅｔＳＯＳ５６の操作と、ポート３４Ｐ及びネットワーク３４を介した通信と、ブレイド１４Ａ及び１４Ｂ間のブレイドパス３０とを介した通信を含めたブレイド１４の全ての通信機能をモニタして調整する。ポート３４Ｐ及びネットワーク３４を介した通信のモニタリングと故障検出のために、各ＣＦａｉｌ６６は、ネットワーク４８とブレイド１４のポート３４Ｐとを介して動作するＳＬＩＰインターフェイス（ＳＬＩＰ）６６Ｓを備えており、ＳＬＩＰ６６Ｓは、ブレイド１４に存在し、ネットワーク調整パケット（ＮＣＰａｃｋ）６６Ｐを反対側のピアブレイド１４とやりとりする。ＮＣＰＡｃｋ６６Ｐは、例えば、ネットワーク調整情報及び通知を備え、ＣＭｏｎｉｔｏｒ６６Ｃによって故障したポート３４Ｐを検出及び識別するために用いられる。特に、各ＳＬＩＰ６６Ｓは、ブレイド１４間の各ネットワーク３４パスを

介して、定期的に、反対側のピアブレイド１４のＳＬＩＰ６６Ｓ及びＣＭｏｎｉｔｏｒ６６ＣにビーコンＮＣＰａｃｋ６６Ｐを送信する。ブレイド１４のＣＭｏｎｉｔｏｒ６６Ｃが、所定の故障検出間隔で、パスを介して反対側のピアブレイド１４からビーコンＮＣＰａｃｋ６６Ｐを受領しない場合、ブレイド１４間のネットワーク３４パスが、故障したものとして検出される。そして、反対側のブレイド１４のポート３４Ｐインターフェイスに故障が起こったと想定される。所定故障検出間隔は、ＮＣＰａｃｋ６６Ｐ通信間の間隔より長く、通常ＣＩＦＳクライアントタイムアウト間隔より短い。本実装においては、この間隔は、１５秒のＣＩＦＳタイムアウト間隔に対し、ほぼ５秒に設定される。

【００５７】図３に示すように、各ＣＦａｉｌ６６は、ＣＭｏｎｉｔｏｒ６６Ｃに応答して任意のＡＲＰ応答６６Ｒを生成するＡＲＰ応答ジェネレータ（ＡＲＰＧｅｎ）６６Ｇと、ネットワーク４８によるクライアント３４Ｃ通信のリダイレクションを管理するために、ＣＦａｉｌ６６の操作にしたがってＣＲＴ４８Ａに存在するＣＲＥ４８Ｅの内容を管理するパスマネージャ（ＰＭ）６６Ｍとを含んでいる。ブレイド１４のＣＭｏｎｉｔｏｒ６６Ｃが、ポート３４Ｐインターフェイスの故障のような、ピアブレイド１４の通信パス故障を判断すると、その情報はＡＲＰＧｅｎ６６Ｇに引き渡され、ＡＲＰＧｅｎ６６Ｇは、クライアント３４Ｃの故障箇所に割り当てられた、あるいは関連するネットワークアドレスを識別するためにＡＲＰテーブル６６Ｔに保存された情報を使用して、故障に関係するポート３４Ｐから接続されたクライアントへの、任意の対応ＡＲＰ応答６６Ｒを生成する。ＡＲＰ応答６６Ｒは、目標となるクライアント３４ＣのＡＲＰテーブルの情報の修正または書き換えを行い、クライアント３４Ｃを対応するポート３４Ｐの動作しているポート３４Ｐ、すなわち、ＡＲＰ応答６６Ｒを生成しているＣＦａｉｌ６６のポート３４Ｐにリダイレクトする。より具体的には、ＡＲＰＧｅｎ６６Ｇにより送信された任意のＡＲＰ応答６６Ｒは、各クライアント３４Ｃに存在するＡＲＰテーブルの修正または書き換えを行い、クライアント３４Ｃからの通信を、ＡＲＰ応答６６Ｒを送信するＡＲＰＧｅｎ６６Ｇを含むブレイド１４の対応するポート３４Ｐに向けようとする。各ＣＦａｉｌ６６は、それにより、故障した通信パスのクライアント３４ＣをＣＦａｉｌ６６が存在するブレイド１４の対応するポート３４Ｐにリダイレクトしようとし、その結果、以下に記述するように、故障したポート３４Ｐと通信するクライアントを機能しているポート３４Ｐを備えたブレイド１４の機能している対応ポート３４Ｐにリダイレクトする。

【００５８】さらに、各ブレイド１４のＰＭ６６Ｍは、ＣＭｏｎｉｔｏｒ６６Ｃの操作と、ＡＰＲＰＧｅｎ６６Ｇによる１つ以上のＡＲＰ応答６６Ｒの生成とに、ＡＲＰ応答６６Ｒの目標であるクライアント３４Ｃに対応するＣＲＴ４８ＡのＣＲＥ４８Ｅを修正することにより応じる。特に、ＰＭ６６Ｍは、故障したエントリ（ＦＥ）４８ＦをＡＲＰ応答が向けられていた各クライアント３４Ｃに対応するＣＲＥ４８Ｅに書き込んで、対応するクライアント４８Ｃの通信がリダイレクトされたことを示し、ＣＲＴ４８Ａにパススルーフィールド（ＰＦ）４８Ｐを設定して、ブレイド１４が１つのモードで動作していることを各ネットワーク４８に知らせる。

【００５９】この後、それ自身のポート３４Ｐを介して、ピアブレイド１４、すなわち、ピアブレイド１４上でサポートされるクライアントファイルシステムあてのクライアント３４Ｃからの通信が受領されると、ネットワーク４８はＰＦ４８Ｐをチェックしてパススルーモード操作が有効であるかどうか判断する。パススルーモードが有効である場合、ネットワーク４８は、ブレイド１４のＢＥＰ４４間のブレイドバス３０パスからなるパススルーパスを介してピアブレイド１４に通信を向ける。さらに、先に記述したリダイレクションの結果として、ネットワーク４８は、ブレイド１４のポート３４Ｐあてのブレイドバス３０パススルーパスを介した通信ではあっても、他方のブレイド１４を通るリダイレクションによりブレイドバス３０パススルーパスを介してリダイレクトされた通信を受領できる。このような場合、ＣＭｏｎｉｔｏｒ６６Ｃ及びＰＭ６６Ｍは、通信ソースであったクライアント３４Ｃに対応するＣＲＥ４８Ｅを修正することで、ネットワーク４８による通信の受領に応じ、ブレイドバス３０パススルーパス及びピアブレイド１４を介してクライアント３４Ｃに通信をルーティングする。これにより、影響を受けたクライアント３４Ｃへの及びそれからのパスの両方向において通信のリダイレクションが完了する。

【００６０】ＨＡＮファイルサーバ１０の別の実装において、各クライアント３４Ｃは、複数のネットワーク３４を介してＨＡＮファイルサーバ１０のポート３４Ｐに接続されることができ、ネットワーク３４は、ローカルエリアネットワーク（ＬＡＮ）、広域ネットワーク（ＷＡＮ）、直接プロセッサ接続またはバス、ファイバオプティックリンク、あるいは前記の組み合わせなどの異なる技術を使用することができることを上述した。これらの実装において、ＣＦａｉｌ６６機構は、ネットワーク３４通信の故障が検出されると上述のように動作するが、さらに、生き残ったブレイド１４にクライアント３４Ｃ通信をリダイレクトするのと同様に、クライアント３４ＣとポートＰ３４Ｐが故障したブレイド１４との間の利用可能及び機能している代わりのネットワーク３４パスを選択してもよい。この実装において、ＣＦａｉｌ６６機構は、上述のように、クライアント３４Ｃ　ＡＲＰテーブル及びＣＲＥ４８Ｅを修正してクライアント３４Ｃ通信をリダイレクトするが、代わりのパスを選択する

際に付加的なオプションを選択する。

【００６１】上述のＩＰパススルー操作に関して、ＨＡＮファイルサーバ１０のＣＦａｉｌ６６機構が、ネットワーク３４とブレイド１４との間の接続場所または原因を識別しようとしないことに注目すべきである。その代わりに、各ＣＦａｉｌ６６は、反対側のブレイド１４のポート３４Ｐインターフェイスに故障が起きたと想定し、ＩＰパススルー操作を開始する。その結果、所定の通信パスのためのＩＰパススルー操作が、ブレイド１４Ａ及び１４Ｂによって同時に実行される。しかしながら、ブレイド１４Ａ及び１４Ｂによって同時に実行されるＩＰパススルー操作は、本発明においては衝突しない。すなわち、例えば、パススルー操作が、ブレイド１４Ａ及び１４Ｂの一方のポート３４Ｐインターフェイスの故障、あるいはブレイド１４Ａ及び１４Ｂの一方へのネットワーク３４リンクの故障の結果である場合、故障に関連するブレイド１４のＣＦａｉｌ６６は、そのポート３４Ｐあるいはネットワーク３４リンクを介して接続されるクライアント３４ＣにＡＲＰ応答６６Ｒを伝達することができない。その結果、故障に関連するブレイド１４のＣＦａｉｌ６６は、そのブレイド１４に対応するクライアント３４Ｃトラフィックをリダイレクトすることができない。しかしながら、反対側のブレイド１４、すなわち、故障に関連しないブレイド１４のＣＦａｉｌ６６は、故障したパスに関連したクライアント３４ＣにＡＲＰ応答６６Ｒを送信し、その結果、ブレイド１４に対応するクライアント３４Ｃトラフィックをリダイレクトことに成功する。ネットワークのパーティションから生じる故障の際には、以下に記述するように、両方のポート３４ＰインターフェイスがブレイドＡ１４及び１４Ｂ間のブレイドパス３０通信パスを介してネットワークパーティションを「橋渡し」できる。その結果、全てのクライアント３４ＣがブレイドＡ１４及び１４Ｂのどちらかと通信できる。

【００６２】最後に、ブレイド１４Ａ及び１４Ｂのどちらかが完全に故障した際には、他方のブレイド１４の生き残った対応ポート３４Ｐにより、故障したポート３４Ｐのサービスの引き継ぎに関して上述した方法で、ＣＦａｉｌ６６を介してＩＰパススルー操作が実行される。ただし、故障したブレイド１４のポート３４Ｐ全てのネットワークサービスは、生き残ったブレイド１４の対応ポート３４Ｐによって引き継がれる。しかしながら、一方のブレイド１４が完全に故障してしまうと、故障したブレイド１４により提供されていたクライアントのＴＣＰ接続が断ち切られてしまうので、ＩＰパススルーの完了後再構築されなければならないことが当業者には明らかであるだろう。その後、故障したブレイド１４上で利用可能だったサービスが生き残ったブレイド１４上で利用可能になり、故障したブレイド１４のクライアントは生き残ったブレイド１４に対してＴＣＰ接続を再構築できる。

【００６３】最後に、上述したＩＰパススルー機構の操作に関して、ＨＡＮファイルサーバ１０によってサポートされるネットワーク３４関連通信操作には、上述したポイントツーポイント、またはクライアント３４ＣからＨＡＮファイルサーバ１０への通信と同様に、例えば、ネットワーク４８のＮｅｔＢＩＯＳ機構により、必要に応じてブロードキャスト通信が含まれることが理解されることと思う。当業者には明らかであるように、ブロードキャスト通信は、特定の受け手へというより複数の受け手にあてられる点でポイントツーポイント通信とは異なるが、ブレイド１４がパススルーモードで動作している時には、クライアント３４Ｃ通信に似た方法で管理される。この場合、ブロードキャスト通信を受けるネットワーク４８は、上述のように、ブレイドがパススルーモードで動作しているかどうかを調べ、もしそうであるならば、ブレイドパス３０パススルーパスを介して反対側のブレイド１４のネットワーク４８に各ブロードキャスト通信を転送する。その結果、その通信は、他のネットワーク４８により直接受けたブロードキャスト通信と同様に取り扱われる。

【００６４】上記に関して、業界標準ＣＩＦＳ仕様書にはクライアントシステム上で動作しているアプリケーションが接続を失った場合の影響が記載、あるいは特定されていないことが当業者にはよく知られている。経験及び実験及びアプリケーション説明書によれば、アプリケーションのＴＣＰ接続が失われた場合の影響はアプリケーションに依存しており、それぞれが故障に対して異なる処理を行う。例えば、あるアプリケーションは、クライアントにＴＣＰ接続を使用する操作を再実行するように指示し、いくつかのアプリケーションは自動的に操作を再実行する。別のアプリケーションは、ユーザに故障を報告するのみである。このため、本実装のネットワークポートフェイルオーバー機構は、これらの機能を実装するための機能を組み込んでおり、それには、各ポート３４Ｐが複数のアドレスに対応することを可能にする、複数のＩＰアドレスをサポートするためにポート３４Ｐを制御するＮｅｔＤＤ６０の機能と、故障したブレイド１４からのＩＰアドレスを転送し、生き残ったブレイド１４上のＩＰアドレスを作成するために必要な機能とが含まれる。ネットワークポートフェイルオーバー機構にはまた、任意のＡＲＰ応答６６Ｒを生成して故障したポート３４Ｐに接続されたクライアントに送信し、さらにクライアントのＡＲＰテーブルのＩＰアドレスが新しいポート３４Ｐをポイントするように変更したり、他のサブシステムの可用性及び故障モニタリング機能と接続してブレイド１４の完全な故障がいつ起きたかを知ったり、故障したブレイド１４リソース名のためのＮｅｔＢＩＯＳ名の変更を行ったりする上述した機能が含まれる。

【００６５】よって、ＨＡＮファイルサーバ１０のＣＦａｉｌ６６機構が、ブレイド１４Ａ及び１４Ｂのポート３４Ｐインターフェイス内のサブネットワークレベルをも含めたどのネットワークレベルに故障が起きても、クライアント３４ＣとＨＡＮファイルサーバ１０のブレイド１４との間の通信を維持あるいは回復できることは明らかである。唯一の必要条件は、ブレイド１４Ａあるいは１４Ｂの少なくとも一方で、１つのネットワーク通信パス及びネットワークインターフェイスが各ネットワーク３４のために機能することである。従って、本発明のＣＦａｉｌ６６機構は、従来技術に典型的な、ネットワーク通信故障のソースと原因とを識別し隔離するのに必要とされる複雑な機構や手順を必要とせず、その一方でまた、衝突する可能性のある故障管理操作を調節し、同期させ、管理するのに必要とされる、これもまた従来技術に典型的な複雑な機構や操作を必要としない。

【００６６】ｂ．ブレイド１４／ブレイド１４通信及び故障処理機構

ＨＡＮファイルサーバ１０の通信機構の中間層が、ブレイドバス３０及びメッセージ４２のような、制御／プロセッササブシステム１４ドメインのブレイド１４Ａ及び１４Ｂドメイン間及びその内部の通信をサポートする通信機構を含むことを上述した。例えば、前述のように、ブレイドバス３０パス及びメッセージ４２は、ブレイド１４間の一連のＨＡＮファイルサーバ１０管理運営通信のために、通信引き継ぎ操作の際のファイルトランザクション操作パスのセグメントとして、ＣＭｉｒｒｏｒ５４Ｍ及びＬＭｉｒｒｏｒ５４Ｌ操作においても使用される。

【００６７】上述し及び図２に示すように、ブレイド１４間のブレイドバス３０通信パスは、ブレイドバス３０、及び、各ブレイド１４のＢＥＰ４４Ｂに存在するＢＥ　ＢｕｓＳｙｓ３８Ｏから構成され、ＢＥ　ＢｕｓＳｙｓ３８Ｏには、Ｓｅｒ－Ｄｅｓ３８Ｒ、ＦＣＣ３８Ｑ、Ｐ－Ｐブリッジ３８Ｈ、ＭＣＨ３８Ｃ、プロセッサ３６Ａなどのエレメントが含まれる。図２には示されていないものの、ＢＥ　ＢｕｓＳｙｓ３８Ｏはまた、プロセッサ３６Ａで、すなわち、ＢＥＰ４４Ｂで動作するＢＥ　ＢｕｓＳｙｓ３８Ｏ制御通信機構を備えている。ＢＥ　ＢｕｓＳｙｓ３８Ｏ制御通信機構は、通常、当業者には明らかな方法で動作し、ＢＥ　ＢｕｓＳｙｓ３８Ｏ及びブレイドバス３０を介する通信操作を実行する。プロセッサ３６Ａ及び３６Ｂ、すなわち、各ブレイド１４のＦＥＰ４４Ｆ及びＢＥＰ４４Ｂはまた、図２あるいは３に示されていないメッセージ４２制御通信機構を実行することが理解されるだろう。メッセージ４２制御通信機構は、通常、当業者には明らかな方法で動作し、メッセージ４２を介する通信操作を実行する。

【００６８】ＢＥＰ４４Ｂ及びＦＥＰ４４Ａ間の通信を提供するメッセージ４２は、各ブレイド１４のメモリ３８Ａの共有メッセージ通信空間と、プロセッサ３６Ａ及び３６Ｂで動作するメッセージング機構とから構成される。メッセージング機構は、通常、当業者には明らかな方法で動作し、メッセージ４２を介する通信操作を実行する。

【００６９】図３に示すように、ＣＦａｉｌ６６には、ＳＬＩＰ６６Ｓ、ＣＭｏｎｉｔｏｒ６６Ｃ、ＡＲＰＧｅｎ６６Ｇとは別の独立した故障処理機構が含まれる。ＳＬＩＰ６６Ｓ、ＣＭｏｎｉｔｏｒ６６Ｃ、ＡＲＰＧｅｎ６６Ｇは、制御／プロセッササブシステム１４ドメインのブレイド１４Ａ及び１４Ｂドメイン間及びその内部の通信についての故障処理のために、制御／プロセッササブシステム１４ドメインへの及びそれからの通信と関連して機能する。図からわかるように、ＣＦａｉｌ６６の相互ブレイド１４ドメイン通信故障処理機構には、ブレイドバス３０及びブレイド１４のＢＥ　ＢｕｓＳｙｓ３８Ｏを含めた、ブレイド１４Ａ及び１４Ｂ間のブレイドバス３０通信リンクの操作をモニタするブレイド通信モニタ（ＢＭｏｎｉｔｏｒ）６６Ｂと、ブレイド１４のメッセージ４２の操作とが含まれる。しかしながら、この接続は図３には示されていない。まずブレイドバス３０を取り上げると、ブレイド１４間、すなわち、ブレイドバス３０あるいはＢＥ　ＢｕｓＳｙｓ３８Ｏのブレイドバス３０通信パスが何らかの理由で故障すると、この故障はＢＭｏｎｉｔｏｒ６６Ｂによって検出され、通常、プロセッサ３６Ａで動作するＢＥ　ＢｕｓＳｙｓ３８Ｏ制御機構が、ブレイドバス３０パスを介して試みられた通信が受領確認されていないと通知する。

【００７０】ブレイドバス３０通信パスの故障の際には、ＢＭｏｎｉｔｏｒ６６Ｂは、ブレイド１４Ａ及び１４Ｂ間の利用可能な通信ルーティングパスに関する情報を保存しているブレイドルーティングテーブル（ＢＲＴ）４８Ｂを読み取る。そこに保存されたパス情報は、例えば、ブレイドバス３０を介する通信のルーティング情報を含み、さらに、ブレイド１４Ａ及び１４Ｂ間の利用可能なネットワーク３４パスのルーティング情報も含む。ＢＲＴ４８ＢはＣＦａｉｌ６６に関連して保存されるが、図３に示すように、本実施例のブレイド１４においては、ＢＲＴ４８Ｂはネットワーク４８と関連して存在する。そのため、ネットワーク３４に関連するルーティングパス情報はすぐに利用されることができ、ＣＲＴ４８Ａの構築などのネットワーク４８の正常動作時にはネットワーク４８にアクセスすることができる。ＢＭＯＮＩＴＯＲ６６Ｂは、故障したブレイドバス３０のパスを除いて、ブレイド１４間の利用可能な通信パスについてのルーティング情報を読み取り、ブレイドバス３０パスの後継あるいは代理で使用される、ブレイド１４のネットワーク４８間の利用可能なネットワーク３４パスを選択する。この関係で、ＢＭＯＮＩＴＯＲ６６Ｂが、ＰＭ６６ＭがＣＲＴ４８ＡのＣＲＥ４８Ｅを修正するのと

同様かつ同時に、全てのＩＰパススルー操作の間にＢＲＴ４８Ｂの内容を修正して、ブレイド１４間の機能していないネットワーク３４パスを示すことに注意しなければならない。この結果、ブレイドバス３０パスの後継パスは、機能しているネットワーク３４パスのみから選択される。

【００７１】ＢＭｏｎｉｔｏｒ６６Ｂは、その後、ＦＥＰ４４Ｆ及びＢＥＰ４４Ｂで動作するＢＥ　ＢｕｓＳｙｓ３８０及びメッセージ４２制御通信機構に、ブレイドバス３０パスにルーティングされる全ての通信を、ＢＥＰ４４Ｂにより直接、あるいはＦＥＰ４４Ｆによりメッセージ４２を介して間接的に、ネットワーク４８及びＰＭ６６Ｍにより選択されたネットワーク３４パスへリダイレクトするという通知を出す。

【００７２】従って、どんな理由によりブレイド１４間のブレイドバス３０通信パスに故障が起きても、ＣＦａｉｌ６６のＣＭｏｎｉｔｏｒ６６Ｃ及びＢＭｏｎｉｔｏｒ６６Ｂ機構は、ネットワーク３４を介してブレイド１４からブレイド１４への通信のために代わりの通信パスを見つけて使用できる。この関係で、ＣＦａｉｌ６６機構が、故障の場所あるいは原因を識別しようとしないので、故障のソースを識別して隔離するのに通常必要となる複雑な機構及び手続と、衝突する可能性のある故障管理操作を調整し、同期させ、管理するのに通常必要となる複雑な機構及び操作とを必要としないことに再び注目すべきである。

【００７３】また、ＨＡＮファイルサーバ１０の通信故障処理機構は、互いに別個に独立して動作するが、これによりまた、衝突する可能性のある故障管理操作を調節し、同期させ、管理するための複雑な機構及び操作を利用する必要がなく、複数の故障ソースあるいは複数の故障を協調して処理できることに注目しなければならない。例えば、ＣＦａｉｌ６６ネットワーク３４故障機構、すなわち、ＣＭｏｎｉｔｏｒ６６Ｃ関連機構によって実行される操作は、ＣＦａｉｌ６６ブレイドバス３０故障機構、すなわち、ＢＭｏｎｉｔｏｒ６６Ｂ関連機構によって実行される操作とは別に実行されるが、クライアント３４Ｃ及びブレイド１４間、そしてブレイド１４間の通信を維持するために機能的に協調して実行される。ブレイド１４間の、そして各クライアント３４Ｃへのネットワーク３４パスが、ブレイドバス３０パスが故障を起こした時に、１つでも機能していれば、通信は、故障のソースあるいは故障の順番に関わらず維持される。

【００７４】例を示すと、第一ブレイド１４と関連するネットワーク３４に故障が起きると、上述のように、第二ブレイド１４を介しての、そしてＣＦａｉｌ６６ネットワーク３４故障機構によりブレイド１４間のブレイドバス３０リンクを介しての第一ブレイド１４への、クライアント３４Ｃ通信のリダイレクションが生じる。次にブレイドバス３０リンクに故障が起こると、ＣＦａｉｌ６６ブレイドバス３０故障機構により、第二及び第一ブレイド１４間で機能している代わりのネットワーク３４パスを介して、第二ブレイド１４及びブレイドバス３０リンクを介してリダイレクトされたクライアント３４通信が再び、第二ブレイド１４から第一ブレイド１４へリダイレクトされる。

【００７５】さらなる例では、第一の故障がブレイドバス３０リンクで起きた場合、ブレイド１４間の通信は、上述のように、ＣＦａｉｌ６６ブレイドバス３０故障機構により、ネットワーク３４を介してブレイド１４間で機能している代わりのパスへリダイレクトされる。この代わりのネットワーク３４パスにおいて次なる故障が起きた場合、この故障はネットワーク３４関連の故障として検出され、ブレイド１４のＣＦａｉｌ６６ネットワーク３４故障機構は、まず、ブレイドバス３０リンクを介してブレイド１４間の先にリダイレクトされた通信をルーティングしようとする。しかしながら、ＣＦａｉｌ６６ブレイドバス３０故障機構は、ブレイドバス３０リンクが機能していないために、ブレイド１４間の利用可能で機能している代わりのネットワーク３４パスを介して先にリダイレクトされた通信をリダイレクトする。

【００７６】従って、ネットワーク３４及びブレイドバス３０の故障がどんな組み合わせあるいは順番で起こっても、クライアント３４Ｃとブレイド１４との間、そしてブレイド１４間の通信を維持するために、ＣＦａｉｌ６６ネットワーク３４及びブレイドバス３０故障機構がさまざまな組み合わせ及び順番で別個の独立した操作を実行することが明らかであろう。また、ブレイドバス３０パスに故障が起きた際に、ブレイド１４間、そして各クライアントへのネットワーク３４パスがたった１つでも機能している限り、故障のソースあるいは故障の順番に関係なく通信は維持される。

【００７７】最後に、この関係で、ブレイド１４のＦＥＰ４４Ｆ及びＢＥＰ４４Ｂ間のメッセージ４２リンクに故障が起きる可能性があることに注意しなければならない。多くの場合、これはブレイド１４が完全に故障した結果であるが、幾つかの場合において、故障はメッセージ４２機構に限定されることができる。メッセージ４２機構に限定された故障の場合、故障が起きたブレイド１４のＦＥＰ４４Ｆは、ブレイド１４のＢＥＰ４４Ｂと、あるいは反対側のブレイド１４と通信することができなくなり、ＢＥＰ４４ＢはブレイドのＦＥＰ４４Ｂと通信できなくなるが、ブレイド１４間のブレイドバス３０リンクを介して反対側のブレイド１４のＢＥＰ４４Ｂ及びＦＥＰ４４Ｆと通信できる。

【００７８】従って、本発明のさらなる実装においては、メッセージ４２に故障が起きたブレイド１４のＢＭｏｎｉｔｏｒ６６Ｂは、ＦＥＰ４４Ｆに関連してブレイドバス３０の明らかな故障を検出するが、ＢＥＰ４４Ｂ

に関連するブレイドバス３０の故障を検出しない。従って、このブレイド１４のBMonitor６６B及びCMonitor６６C機構は、PM６６Mによって選択されたネットワーク３４パスを介して、FEP４４Pから全ての通信をBEP４４Bへ、あるいは反対側のブレイド１４へリダイレクトし、BEP４４BからFEP４４Fへの全ての通信をブレイドバス３０、及びFEP４４Fのために選択されたネットワーク３４パスを介するルートへリダイレクトするが、ブレイドバス３０を介するBEP４４B通信をリダイレクトしない。

【００７９】故障が起きなかったブレイド１４においては、BMonitor６６B機構は、メッセージ４２が故障したブレイド１４のFEP４４Pへの通信について明らかなブレイドバス３０パス故障を検出するが、そのブレイド１４のBEP４４Bへの通信についてのブレイドバス３０パス故障を検出しない。従って、このブレイド１４のBMonitor６６B及びCMonitor６６C機構は、反対側のブレイド１４のFEP４４Fあての全ての通信を、上述のように、代わりのネットワーク３４パスを介してリダイレクトするが、反対側のブレイド１４のBEP４４Bあての通信をリダイレクトしない。

【００８０】c．記憶サブシステム１２／ブレイド１４故障処理機構

上述のように、HANファイルサーバ１０の故障処理機構の最下層レベルには、記憶サブシステム１２の通信パス構造及びRAID４６によって提供されるRAIDF４６F機構とが含まれる。RAIDファイル機能は、当業者にはよく知られているため、ここでは本発明を理解するのに必要な場合のみ説明し、以下には、記憶サブシステム１２内部の、そしてサブシステム１２及びブレイド１４間の通信パスに焦点を当てて説明する。

【００８１】図１に示すように、そして上述したように、記憶サブシステム１２には複数のハードディスクドライブ１８から構成されるドライブバンク１６が含まれる。各ハードディスクドライブ１８は、デュアル記憶ループモジュール２０A及び２０Bを介して双方向に読み取り／書き込みアクセスされる。記憶ループモジュール２０A及び２０Bそれぞれには、MUXBANK２２A及び２２Bが含まれ、各MUXBANK２２には、複数のMUX２４とループコントローラ２６A及び２６Bとが含まれる。各ループコントローラモジュール２０のMUX２４とループコントローラ２６とは、MUXループバス２８A及び２８Bを介して双方向に相互接続されている。図からわかるように、MUXBANK２２A及び２２Bそれぞれには、対応するディスクドライブ１８の１つに対応して接続されるMUX２４Dが含まれる。そのため、ドライブバンク１６の各ディスクドライブ１８は、MUXBANK２２A及び２２Bそれぞれの対応するMUX２４Dに接続されて双方向に読み取り／書き込みされる。MUXBANK２２A及び２２Bそれぞれには、さらに、MUX２４CA及びMUX２４CBを介して対応するコンピュートブレイド１４A及び１４Bの一方が双方向に接続されており、コンピュートブレイド１４A及び１４Bは、ブレイドバス３０を介して双方向に接続されている。

【００８２】従って、各ディスクドライブ１８は、MUXバンク２２AのMUX２４DとMUXバンク２２BのMUX２４Dとに双方向に接続されている。MUXバンク２２AのMUX２４は、ループバス２６Aを介して相互接続されている一方、MUXバンク２２BのMUX２４は、ループバス２６Bを介して接続されている。そのため、各ディスクドライブ１８は、ループバス２６A及びループバス２６B両方を介してアクセス可能である。さらに、プロセッサブレイド１４Aは、ループバス２６Aと双方向に通信する一方、プロセッサブレイド１４Bは、ループバス２６Bと双方向に通信し、プロセッサブレイド１４A及び１４Bは、ブレイドループ（ブレイド）バス３０を介して直接相互接続されて通信する。

【００８３】従って、記憶サブシステム１２内部の下層レベルの通信故障処理機構が、基本的に、各ディスクドライブ１８とプロセッサブレイド１４A及び１４Bとの間に複数の予備のアクセスパスを提供する受動的なパス構造であることがわかるだろう。このため、プロセッサブレイド１４A及び１４Bは、記憶サブシステム１２内部の１つ以上の通信パスで故障が起きた際には、対応するループバス２６を介して直接、あるいは他方のプロセッサブレイド１４を介して間接的に、ディスクドライブ１８のどれとでも双方向通信が可能であり、互いに直接通信できる。１つ以上のディスクドライブ１８内で起きる故障のための故障処理機構は、上述のRAIDF４８F機構から構成される。

【００８４】また、記憶サブシステム１２の受動パス構造が、通信機構と、ブレイド１４のCFail６６ネットワーク３４及びブレイドバス３０故障機構とは別々に独立して動作するものの、クライアント３４Cと、クライアント３４のファイルシステムが存在するディスクドライブ１８との間の通信を保証するために、ブレイド１４の機構と協調して動作することがわかるだろう。また、これらの機構は、複雑な故障検出、識別、隔離機構の利用と、複雑な故障管理調整、同期、管理機構の利用とを廃して、高レベルのファイルシステム可用性を提供する。

【００８５】５．HANファイルサーバ１０のファイルトランザクション故障処理機構とHANファイルサーバ１０の通信故障処理機構の相互運用（図１、２、３）

本実施例のHANファイルサーバ１０が、多数の高可用性機構、すなわち、HANファイルサーバ１０の１つ以上のコンポーネントに故障が起きた際にも、HANファイルサーバ１０がクライアントへのファイルサーバサー

ビスを中断せずに提供し続けることを可能にする機構を備えることを上述した。これらの機構の多くは、基本ＲＡＩＤＦ４６Ｆ機能のように、従来技術の代表的なものであり、当業者にとっては周知のものである。そのため、本発明に関係しない限り詳細な説明を省く。

【００８６】しかしながら、一般的には、ＨＡＮファイルサーバ１０のコンポーネントに故障が起きた際には、ＨＡＮファイルサーバ１０の生き残ったコンポーネントが、高可用性機構の操作により、故障したコンポーネントによって実行されていたタスク及びサービスを引き継ぎ、これらのサービスの提供を続ける。このような高可用性機構の操作には数多くの機能があり、そのような機構がこれらの機能を達成するためには幾つかの操作を実行する必要があることが当業者には明らかであろう。例えば、高可用性機構は、コンポーネントの故障を識別し、故障したコンポーネントから生き残ったコンポーネントへソースあるいは機能の引き渡しあるいは移転を行い、故障したコンポーネントによって提供されていたサービス及び機能が外からわかるように中断されないように生き残ったコンポーネントに引き継がれたリソースの状態を回復し、故障したコンポーネントの置換あるいは訂正を行ない、修復後には故障していたコンポーネントにリソースを引き渡すあるいは移動する必要がある。

【００８７】通信に関して上述したように、ＨＡＮファイルサーバ１０のファイルトランザクション及び通信機構は、独立して動作する。そして以下にさらに詳細に説明されるように、本発明のＨＡＮファイルサーバ１０の高可用性機構は、ＨＡＮファイルサーバ１０の多数の異なる機能レベルで動作する。通常、異なるグループ、あるいは異なるタイプの操作及び機能は、ＨＡＮファイルサーバ１０の各機能レベルで実行される。従って、高可用性機構はそれぞれ異なり、各レベルで、そしてシステムとしてのＨＡＮファイルサーバ１０のために、独立しながらも協調して動作して高レベルのサーバ可用性を提供する。以下にさらに詳細にこれらの機構の構造及び操作と、これらの機構の相互運用とを説明する。

【００８８】例えば、ＨＡＮファイルサーバ１０における最上層レベルの機能は、クライアント通信タスク及びサービスを実行する通信レベル、すなわち、クライアントと、ネットワーク３４を介してＨＡＮファイルサーバ１０によってサポートされるクライアントファイルシステムとの間の通信である。この通信レベルの中心機能は、ネットワーク４８の機構とＨＡＮファイルサーバ１０の関連コンポーネントとによって提供される。通信レベルでの高可用性機構には、ＣＦａｉｌ６６のような故障検出機構が含まれ、通信レベルでの故障を処理する多数の異なる機構を提供する。例えば、ブレイド１４Ａ及び１４Ｂのうちの一方で１つ以上のポート３４Ｐを介する通信に故障が起きた場合、ピアブレイド１４のＣＦａｉｌ６６は故障を検出し、ネットワーク４８と連携して、クライアントと故障したポート３４Ｐとの間の全ての通信を、ピアブレイド１４の機能している対応ポート３４Ｐにリダイレクトする。ピアブレイド１４では、その内部のネットワーク４８が、ブレイドバス３０を介して、故障したポート３４Ｐを有するブレイド１４のＪＦｉｌｅ５０に通信をルーティングする。その結果、故障したポート３４Ｐは、ピアブレイド１４のポート３４Ｐと、ブレイドバス３０及びメッセージ４２を介するＦＥＰ４４Ｆ－ＢＥＰ４４Ｐ通信パスからなる相互ブレイド１４通信パスとを介してバイパスされる。この関係で、ブレイド１４の高レベルファイルトランザクション機構について以下の記述により説明されるように、ネットワーク４８の高可用性機構は、高レベルファイルトランザクション機構の高可用性機構を相互運用して、実際の、そして例えば、ブレイド１４ＪＦｉｌｅ５０のあるいはブレイド１４全体の故障から生じる明らかなネットワーク３４関連通信故障に対処する。

【００８９】ブレイド１４における次のレベルの機能は、高レベルファイルトランザクション機能及びサービスから構成される。そこでは、高レベルトランザクション機能の中心機能及び操作は、ＪＦｉｌｅ５０及び関連する高レベルファイル機構により提供される。上述のように、ＨＡＮファイルサーバ１０の高レベルファイル機能レベルでの高可用性機構には、ＣＭｉｒｒｏｒ５４Ｍを備えたＷＣａｃｈｅ５０ＣとＬＭｉｒｒｏｒ５４Ｌを備えたログ５０Ｌとが含まれ、これらの機構は、ブレイド１４内部の高レベルファイル機構の故障を処理する。上述のように、ＷＣａｃｈｅ５０Ｃは、従来方法で動作してデータトランザクションをキャッシュし、ＣＭｉｒｒｏｒ５４Ｍは、ＷＣａｃｈｅ５０Ｃに影響するＦＥＰ４４Ｆに故障が起きた際、ＷＣａｃｈｅ５０Ｃの内容を回復できる。ログ５０Ｌは、ブレイド１４とともに動作してＪＦｉｌｅ５０により実行されるファイルトランザクションの履歴を保存する。これにより、ログ５０Ｌは、例えば、トランザクションが記憶サブシステム１２の固定記憶装置に完全にコミットされる前にファイルトランザクションの損失を生じる、ＪＦｉｌｅ５０あるいは記憶サブシステム１２の故障の際、失われたファイルトランザクションを再実行及び回復させることができる。

【００９０】しかしながら、ＬＭｉｒｒｏｒ５４Ｌ機構は、ＬＭｉｒｒｏｒ５４Ｌがミラーリングするログ５０Ｌが存在するブレイド１４内部で動作せず、代わりに、ブレイド１４を横断して動作して、各ＬＭｉｒｒｏｒ５４Ｌが、反対側のピアブレイド１４のログ５０Ｌの内容をミラーリングして保存できるようにしている。その結果、ＬＭｉｒｒｏｒ５４Ｌ機構は、反対側のピアブレイド１４に壊滅的な故障が起きた場合にも反対側のピアブレイド１４のログ５０Ｌの内容を保存し、故障していたブレイド１４がサービスを再開した際に、失われたファ

イルトランザクションを故障していたブレイド１４で再実行及び回復することができる。

【００９１】さらに、生き残ったブレイド１４内部に故障したブレイド１４の失われた可能性のあるファイルトランザクションのレジデント履歴を備えることにより、ＬＭｉｒｒｏｒ５４Ｌ機構はまた、生き残ったブレイド１４に故障したブレイド１４によってサポートされていたクライアントのサポートを引き継がせることができることに注目すべきである。すなわち、ネットワーク４８機構について上述したように、生き残ったブレイド１４のネットワーク４８及びＪＦｉｌｅ５０は、故障したブレイド１４のクライアントを生き残ったブレイド１４にリダイレクトすることにより、故障したブレイド１４によって先にサポートされていたクライアントのサービスを引き継ぐ。この処理では、上述のように、生き残ったブレイド１４のネットワーク４８機構は、生き残ったブレイド１４のＪＦｉｌｅ５０に、引き継がれたＩＰアドレスあてのデータトランザクションを向けることにより、故障したブレイド１４のＩＰアドレスを引き継ぐ。生き残ったブレイド１４のＪＦｉｌｅ５０は、生き残ったブレイド１４がローカルファイルシステムを備えるという仮定の下に、新しいクライアントとして故障したブレイド１４のクライアントを引き継ぎ、その後は、引き継がれたクライアントを自分のクライアントとしてサービスを行う。そのサービスには、引き継がれたデータトランザクションを処理することと並行して全ての引き継がれたデータトランザクションを記録することが含まれる。生き残ったブレイド１４は、ローカルリカバリログ、すなわち、生き残ったブレイド１４に存在するＬＭｉｒｒｏｒ５４Ｌを使って引き継いだＩＰアドレスのデータトランザクションを記録するとともに、レジデントＬＭｉｒｒｏｒ５４Ｌに保存されたファイルトランザクション履歴を使用して故障したブレイド１４の失われたファイルトランザクションを再実行及び再構成し、故障したブレイド１４のクライアントのファイルシステムを所望の状態に回復することができる。この関係で、生き残ったブレイド１４のＪＦｉｌｅ５０は、故障したブレイド１４に向けられていたファイルトランザクションの初期アドレスを基にしてネットワーク４８からの通知により、あるいはレジデントＬＭｉｒｒｏｒ５４Ｌの内容を調べて保存されたファイルトランザクションと相互に関連する「新しい」クライアントファイルトランザクションがあるかどうか判断することにより、「新しい」クライアントが故障したブレイド１４から移転されたクライアントであるかを判断できる。

【００９２】最後に、ＨＡＮファイルサーバ１０の最下層レベルのファイルトランザクション機能は、ＲＡＩＤ４６によってサポートされるＲＡＩＤ４６ファイルトランザクション機能及びサービスから構成される。ＲＡＩＤＦ４６Ｆ機能は、それ自身、上層レベルの高可用性機構から独立して動作することがわかるだろう。しかしながら、通信レベル及び高レベルファイルトランザクション機構は、例えば、デュアルブレイド１４Ａ及び１４Ｂ、ループバス２６Ａ及び２６Ｂ、ＭＵＸループバス２８Ａ及び２８Ｂを介する代わりの通信パスの提供と連携してＲＡＩＤＦ４６Ｆ機能と協調的に動作し、ディスクドライブ１８へのアクセス可能性を高めていることがわかるだろう。

【００９３】従って、ＨＡＮファイルサーバ１０に設けられた通信レベル及び高レベルファイルトランザクション機構と代わりの通信パスとは、ＲＡＩＤＦ４６Ｆ機能と協力してネットワーククライアントへのファイルシステム共有資源、すなわち、記憶空間の可用性を高めることが上記より理解されることができる。また、ＨＡＮファイルサーバ１０に設けられた通信レベル及び高レベルファイルトランザクション機構と代わりの通信パスとが、複雑な故障検出、識別、隔離機構の利用、及び複雑な故障管理調整、同期、管理機構の利用を廃して、上記の効果を達成することが理解されるだろう。

【００９４】よって、要約すると、数多くの異なる機構が故障したコンポーネントを識別するために用いられ、その機構は、コンポーネントと、コンポーネントが存在するＨＡＮファイルサーバ１０のサブシステムと、コンポーネントの故障によるＨＡＮファイルサーバ１０の操作への影響とに依存して特定されることが上記から理解される。例えば、ＲＡＩＤＭ４６Ｍ機能が、ファンや電源装置のようなコンポーネント、及びブレイド１４Ａ及び１４Ｂの類似のコンポーネントの故障をモニタして検出する一方、ＲＡＩＤＦ４６Ｆ機能は、ディスクドライブ１８のファイルシステム操作のエラー及び故障をモニタ、検出、修正あるいは補正する。ＲＡＩＤ４６機構によってモニタされるコンポーネントの多くは故障が起きても、システムとしてのＨＡＮファイルサーバ１０レベルでのデータの可用性を危うくすることはないが、そのコンポーネントを修復するための処置を取ることができるように管理インターフェースを通じて検出及び連絡されなければならないことがわかるだろう。さらなる例では、ＨＡＮファイルサーバ１０のネットワーク管理機能は、ネットワーク３４の状態と、ＨＡＮファイルサーバ１０のネットワーク３４通信関連コンポーネントとをモニタし、それぞれの故障に適した方法で、ＨＡＮファイルサーバ１０とＨＡＮファイルサーバ１０のクライアントとの間での通信の故障に対応する。ネットワークをモニタするために、ネットワーク管理機能は、ＨＡＮファイルサーバ１０自身のネットワーク通信をテストするためのセルフチェックを生成し、外部ネットワークと通信しているかどうか判断する。例えば、このセルフチェックがネットワークパスのどれかで失敗する場合、故障したネットワークパスによってサポートされていた通信は、上述のように別のネットワークパスに引き継がれ

る。さらに別の例においては、ＲＡＩＤ４６機能がブレイド１４の故障を検出すると、この故障が上述のようにファイルシステム機能に連絡され、その結果、フェイルオーバー処理が適切なファイルシステムレベルで実行されることができる。

【００９５】故障処理過程での次のステップ、すなわち、生き残ったリソースへの故障したリソースの移転は、通常、既知の生き残った場所にリソースを再割り当てすることにより実行される。ネットワーク機能の故障の場合、移転は、上述のように、故障したデバイスの機能を引き継ぐことのできる、先に識別されたネットワークアダプタに対して行われる。故障したのがブレイド１４である場合は、ピアブレイド１４が故障したブレイド１４からファイルシステムを引き継ぐ。

【００９６】故障したコンポーネントから生き残ったコンポーネントへのリソースの移転には、そのリソースが生き残ったコンポーネント上で利用可能にされる前にリソースの動作状態を変更あるいは修正する必要がある。例えば、ネットワークコンポーネントの故障の場合、新しいネットワークアドレスが既存のアダプタに付加されなければならず、ブレイド１４の故障のようにファイルシステムに影響を与える故障の場合には、トランザクションログを再実行して故障で失われたデータを置換する。

【００９７】先に記述したように、ＨＡＮファイルサーバ１０のコンポーネントの多くは、ＨＡＮファイルサーバ１０から取り外して、動作しているコンポーネントに置換することができる、ホットスワップ可能なコンポーネントである。一旦コンポーネントを置換すると、生き残ったコンポーネントにより引き継がれたリソースは初期のコンポーネントに、つまりは、初期のコンポーネントが置換されたものに戻されなくてはならない。従って、上述のような適切なサブシステムの回復機構では、生き残ったコンポーネントに移転されたリソースは置換されたコンポーネントに移行される。この処置は、通常、システムアドミニストレータにより手動で、そしてサービスの中断が受け入れ可能及び処理可能な時に行なわれる。

【００９８】本発明が、ここに例として使われたファイルサーバと同様に、例えば通信サーバ、さまざまなタイプのデータプロセッササーバ、プリンタサーバなどの、クライアントとの信頼できる通信と、データあるいは処理トランザクションの保存及び回復とを必要とするあらゆる形式の共有リソースに実装可能であることが当業者には明らかであろう。また、本発明が、例えば、異なるＲＡＩＤ技術、異なる保存技術、異なる通信技術、そして画像処理などの他の情報処理手法及び技術を使用するファイルサーバの実装にも、同様に適応できるとともに実装可能であることが明らかであろう。異なる形式の共有リソース、異なるリソースマネージャ、異なるシステム構成及びアーキテクチャ、異なるプロトコルにも本発明が適応できることは当業者には明らかであろう。

【００９９】従って、本発明が、実施例の装置及び方法について特に説明され記述されてはいても、ここに説明され、付属の請求項によって規定される本発明の範囲を超えない限り、形式、詳細、実装におけるさまざまな変更、変形、修正を本発明に加えることができることが当業者には明らかであろう。よって、本発明のあらゆる変形及び修正を本発明の範囲内に収まるようにカバーすることが付属の請求項の目的である。

【図面の簡単な説明】

【図１】　本発明が実装されることのできるネットワークファイルサーバのブロック図である。

【図２】　図１のファイルサーバのドメインにおけるプロセッサのコアのブロック図である。

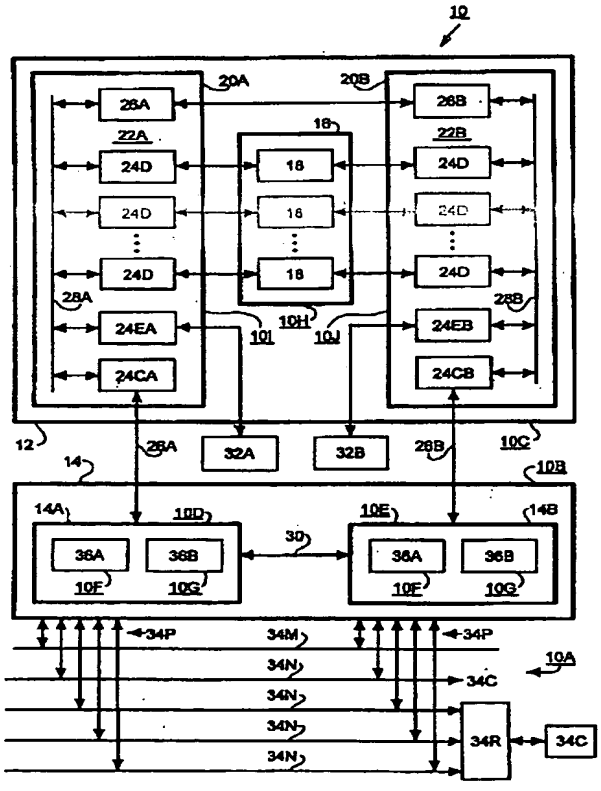
【図３】　図１のファイルサーバのドメインをさらに詳細に示した概略図である。

【符号の説明】

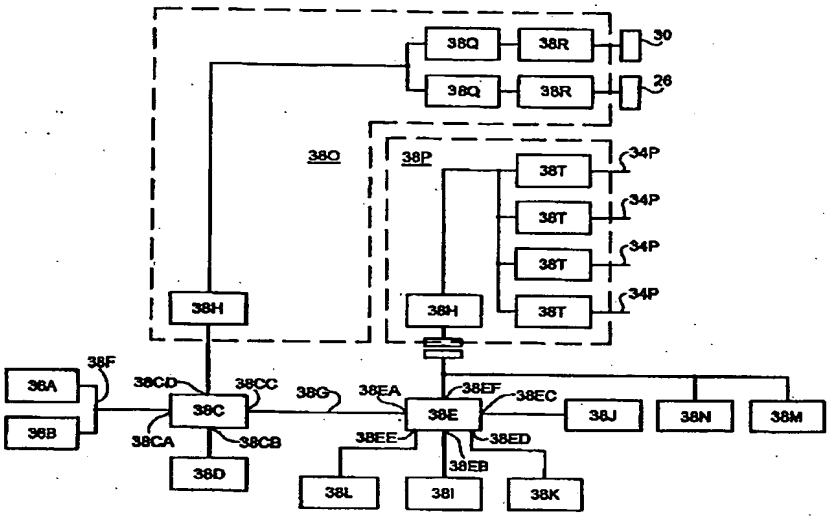
１０	ＨＡＮファイルサーバ
１２	記憶サブシステム
１４	制御／プロセッササブシステム
１４Ａ、１４Ｂ	プロセッサブレイド
１６	ドライブバンク
１８	ディスクドライブ
２０Ａ、２０Ｂ	記憶ループモジュール
２２Ａ、２２Ｂ	マルチプレクサバンク
２６Ａ、２６Ｂ	ループコントローラ
２８Ａ、２８Ｂ	ＭＵＸループパス
３０	ブレイドパス
３２Ａ、３２Ｂ	外部ディスクアレイ
３４Ｃ	クライアント
３４Ｍ	管理ネットワーク
３４Ｎ	クライアントネットワーク
３４Ｐ	ネットワークポート
３４Ｒ	ルータ
３６Ａ、３６Ｂ	処理ユニット
３８Ｃ	メモリコントローラハブ
３８Ｄ	メモリ
３８Ｅ	入出力コントローラハブ
３８Ｆ	フロントサイドバス
３８Ｇ	ハブリンクバス
３８Ｈ	Ｐ－Ｐブリッジ
３８Ｉ	ファームウェアメモリ
３８Ｊ	ハードウェアモニタ
３８Ｋ	ブートドライブ
３８Ｌ	スーパーＩ／Ｏデバイス
３８Ｍ	ＶＧＡデバイス
３８Ｎ	ネットワークデバイス
３８Ｏ	バックエンドバスサブシステム
３８Ｐ	フロントエンドバスサブシステム

３８Ｑ	ファイバチャネルコントローラ		５０Ｌ	トランザクションログ
３８Ｒ	シリアライザ／デシリアライザデバイス		５０Ｍ	ログメモリ
			５４Ｌ	ログミラー機構
３８Ｔ	ネットワークデバイス		５４Ｍ	キャッシュミラー機構
４０	オペレーティングシステム	05	５６	ネットワークスタックオペレーティングシステム
４２	メッセージパッシング機構			
４４Ｂ	バックエンドプロセッサ		５８	ＴＣＰ／ＩＰプロトコルスタック
４４Ｆ	フロントエンドプロセッサ		６０	ネットワークデバイスドライバ
４６	ＲＡＩＤ機構		６２	ＣＩＦＳ
４６Ｍ	ＲＡＩＤモニタ機構	10	６４	ＮＦＳ
４６Ｆ	ＲＡＩＤファイル機構		６６	通信フェイルオーバー機構
４８	ネットワーク機構		６６Ｂ	ブレイド通信モニタ
４８Ａ	クライアントルーティングテーブル		６６Ｃ	通信モニタリング処理／プロトコル機構
４８Ｂ	ブレイドルーティングテーブル			
４８Ｅ	クライアントルーティングエントリ	15	６６Ｇ	ＡＲＰ応答ジェネレータ
４８Ｐ	パススルーフィールド		６６Ｍ	パスマネージャ
５０	ジャーナルファイルシステム		６６Ｐ	ネットワーク調整パケット
５０Ｃ	書き込みキャッシュ		６６Ｒ	ＡＲＰ応答
５０Ｆ	ファイルシステム機構		６６Ｓ	ＳＬＩＰインターフェイス
５０Ｇ	ログジェネレータ	20		

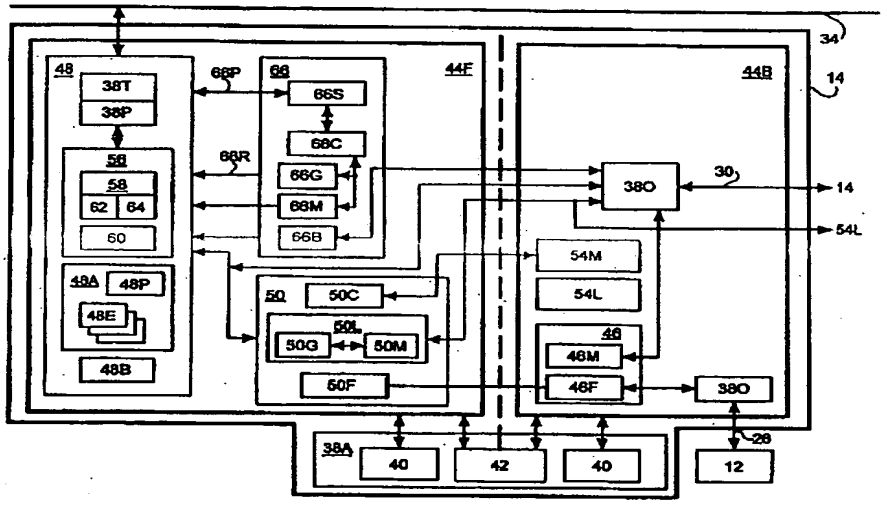
【図１】



【図２】



【図３】



フロントページの続き

(72)発明者　ジェームズ　グレゴリー　ジョーンズ
アメリカ合衆国　ノースカロライナ州　45
27615　ローリー　モントーク　ドライブ
8708

Ｆターム(参考)　5B082 DD00 DE02
5B083 AA08 BB01 CD11 EE11
5B089 GA12 JB17 KA12 KB02 KC15
KG05 KG08 ME02 ME04

(19) JAPANESE PATENT OFFICE

PATENT ABSTRACTS OF JAPAN

(11) Publication number: **2000242434 A**

(43) Date of publication of application: **08.09.00**

(51) Int. Cl **G06F 3/06**

(21) Application number: **11344260**

(22) Date of filing: **03.12.99**

(30) Priority: **22.12.98 JP 10364079**

(71) Applicant: **HITACHI LTD**

(72) Inventor: **MATSUNAMI NAOTO**
OEDA TAKASHI
YAMAMOTO AKIRA
AJIMATSU YASUYUKI
SATO MASAHIKO

(54) **STORAGE DEVICE SYSTEM**

(57) Abstract:

PROBLEM TO BE SOLVED: To construct a storage device system corresponding to the scale or request of a computer system so that the extension of a storage device system and improvement in reliability in the future are easily realized.

SOLUTION: This system 1 has a plurality of subsets 10 having a storage device for holding data and a controller for controlling the storage device and switch devices 20 arranged between the subsets 10 and a host 30. Each switch device 20 has a managing table for holding management information for managing the configuration of the storage device system 1. According to the management information, address information contained in frame information outputted by the host 30 is translated and the frame information is distributed to the subsets 10.